Untersuchungen zur sprachübergreifenden, bilingualen Suche mit Hilfe der Konzeptnetz-Technologie der SENTRAX-Engine

Vom Fachbereich III

– Informations- und Kommunikationswissenschaften –

der Universität Hildesheim

zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr.rer.nat.)
genehmigte
Dissertation

Von
Suriya Na nhongkai

aus Ratschaburi (Thailand)

Berichterstatter: Prof. Dr. Hans-Joachim Bentz
Prof. Dr. Manfred Wettler

eingereicht am: 01. Februar 2006
mündliche Prüfung am: 03. Juli 2006

DANKSAGUNG

Für die freundliche Unterstützung, die geduldige Betreuung und die Überlassung des Themas möchte ich mich bei meinem Doktorvater, Herrn Prof. Dr. Hans-Joachim Bentz, ganz herzlich bedanken. Mein Dank gilt ebenso Herrn Prof. Dr. Manfred Wettler für die bereitwillige Übernahme des Koreferats.

Als Promotionsstipendiat bedanke ich mich sehr beim thailändischen Hochschulministerium. Es hat mir diesen Aufenthalt in Deutschland sowie die intensive und ungestörte Beschäftigung mit den Forschungsaufgaben ermöglicht.

Darüber hinaus gilt der Dank allen Kollegen am Fachbereich Mathematik, Fakultät der Wissenschaft, Kasetsart Universität, Thailand, für ihre geduldige Antizipation.

Das Verfassen der Arbeit wäre nicht so gut gelungen, wenn Herr Dr. Andreas Dierks, Herr Martin Zander, Universität Hildesheim, mir nicht bei den Korrekturen geholfen hätte. Ich bedanke mich für das geduldige Überprüfen meiner deutschen Entwürfe. Gleichermaßen danke ich für die zahlreichen Anregungen von meiner Frau, meinen Eltern und allen Freunden.

Für die Bereitstellung der Hilfsanwendung des TreeTagger-Programms zu wissenschaftlichen Zwecken bin ich dem IMS (Universität Stuttgart), namentlich Herrn Dr. Helmut Schmid, und für die Bereitstellung des Europarl parallelen Korpus Herrn Dr. Philipp Koehn, School of Informatics (University of Edinburgh), sehr dankbar.

KURZFASSUNG

Ein Hindernis bei der Suche nach benötigter Information – speziell bei einer krosslingualen Suche – ist eine ungünstig formulierte Anfrage. Die Wörtervielfalt, aus denen eine Anfrage zusammengesetzt werden kann, verursacht oft eine ungenügende Übereinstimmung mit den Formulierungen im gesuchten Dokument und schmälert die Leistungsfähigkeit der Suche. Wenn man die "Bedeutung" einer Wortsammlung an die Engine übergeben könnte – anstelle isoliert verarbeiteter Worte –, dann könnte eine Wirkung der Suchanfragen erzielt werden, die als gleichmäßiger empfunden würde. Dieser Gedanke wurde bei der Entwicklung einer neuartigen Retrievaltechnologie verfolgt und führte zur sogenannten "Essence Extractor Engine", kurz SENTRAX [SENT04]. Der dahinter liegende Index entsteht aus der Verarbeitung von in den Dokumenten nahe zusammenstehenden, bedeutungstragenden Begriffen (Kookkurrenzen) und erlaubt eine Definition und Übertragung von "Konzepten", die zwar durch Worte ausgedrückt oder beschrieben werden, aber eine gewisse Unabhängigkeit von der spezifischen Wortwahl haben. Diese Technologie stand für die vorliegende Arbeit zur Verfügung und wurde für die Problemstellung des Themas ausgenutzt. Bei der bilingualen Suche kann nämlich die Übertragung eines Konzeptes – statt der wortweisen Übersetzung der Anfrage – die Mehrdeutigkeiten entscheidend vermindern, da das Konzept den assoziierten Zusammenhang mit den übersetzten Begriffe bewahrt und die Verbindung zu den Umgebungen in den Texten herstellt. Diese Wirkung und Auswirkung wird untersucht und dargestellt. Weitere Funktionen der SENTRAX-Engine (z.B. Stringtoleranz von Eingabeworten und Ähnlichkeitsvergleich von Trefferdokumenten) sowie eine grafische Mensch-Maschine-Schnittstelle erweisen sich als günstig für das Vorhaben.

Die nötigen Vorverarbeitungsmethoden werden entworfen, da zwei Indexe für die bilinguale Suche zusammenwirken. Drei wichtige Teile lassen sich nennen: erstens die Vorarbeit, wo die Erstellung des jeweiligen Konzepts geschieht, zweitens die Brücke, die das Suchkonzept der Ausgangsprache zur Zielsprache überträgt, und schließlich ein Konzeptsvergleichmaß, womit das Gleichgewicht des Konzeptes nach der Übertragung kontrolliert wird. Gegenwärtig laufen diese drei Stufen noch nicht vollautomatisch in

der SENTRAX ab, sondern erlauben manuelle Eingriffe. Ungeachtet dieser technischen Unvollständigkeit des Systems lassen sich aber alle Hypothesen nachprüfen.

Die Ergebnisse zeigen, dass die bilinguale Suche mittels Konzeptnetzen sehr leistungsfähig ist. Die Mehrdeutigkeit einer Übersetzung kann durch Betrachtung der Zusammenhänge und der Assoziationen vermindert werden. Es gibt auch einen Lernprozess beim Suchenden und hilft dem Nutzer bei der Entscheidung, ob eine Fortsetzung der begonnenen Suchrichtung Sinn macht oder diese abgebrochen werden muss, weil die benötigte Information vielleicht nicht in der Datenbasis steckt. Durch die grafische Darstellung werden die mit den Suchwörtern assoziierten Begriffe zur Auswahl angeboten und sind in beiden Sprachen (teilweise) vergleichbar. Dadurch kann man sicherstellen, dass die dahinter liegenden Dokumente von den gleichen bzw. ähnlichen Themen handeln.

ABSTRACT

A problem in the search for information is an unsuitable formulated query – in particular with respect to cross language document retrieval. A query can be built in many ways according to the combination of words used. This often causes insufficiency and ineffectiveness in the retrieval process. The idea to maintain the "concept" or "meaning" of a set of terms and process it within the search -instead of a pure list of singular items- led to the development of an innovative retrieval engine, the so called SENTRAX ("essence extraxtor engine" [SENT04]). The underlying index built from the documents refers to collections of meaningful terms that are close neighbours in the texts (cooccurrences). It allows a definition and a processing of concepts which are described by words but have a certain independency from the chosen terms. This technology was extensively used for this thesis. As to the task of the bilingual search the transfer of a concept can strongly reduce the ambiguity which normally comes along with the word by word translation of the query. A concept retains associations of the translated terms as well as it connects to the neighbourhoods in the texts. For theses reasons the bilingual search can be well done by the SENTRAX method. In addition to this some other features of this engine (e.g. error tolerance of strings, similarity clustering of document hits, graphic user interface) have shown to be very useful for this project.

The binding construction units and the necessary pre-processing methods are designed in order to create the bilingual search by two SENTRAX indexes. This works in three steps. First the pre-processing, which is responsible for building a concept. Second is the bridge, which transfers the searching query from the source language to the target language. Finally there is a concept comparison measure, which controls the equilibrium of the concept after its transfer. At present these three parts do not run fully automatic within with the SENTRAX but allow manual control. Despite such incompleteness of the system the hypotheses can be tested.

It can be stated by the results of the examination that the bilingual search can be done very well via a concept network. Ambiguities of the translation can be decreased by the consideration of context connections and of associations. Besides this there is a learning

process while operating on the tasks which supports the user in the decision whether continuing with the search or to stop it, because the necessary information was never contained in the database. The graphical interaction tools offer terms associated with the input, and can be compared (partly) in both languages. By this incident it can be checked if the documents in the database deal with the same or similar topic.

บทคัดย่อ

ปัญหาหนึ่งที่เป็นอุปสรรคต่อการได้มาของข้อมูลที่ต้องการนั้นเกิดจากคำที่ใช้ค้นหาถูกจำลองขึ้นมาอย่างไม่เหมาะสม โดยเฉพาะอย่างยิ่งในการค้นคืนข้ามภาษา คำที่ใช้ในการค้นหาสามารถประกอบขึ้นจากคำหลายๆ คำในหลายๆรูปแบบ การประกอบกันของคำที่ใช้ในการค้นหาที่ไม่เพียงพอ ซึ่งไม่สามารถนำไปสู่ความสอดคล้องกันกับเอกสารที่ค้นหา ส่งผลให้การค้นคืนมีประสิทธิภาพไม่มากพอ ความสมดุลย์ในความหมายของกลุ่มคำที่ใช้ในการค้นหาสามารถรักษาไว้ได้ด้วยคอนเซ็ปที่ถูกประกอบขึ้นจากกลุ่มคำที่เหมาะสม แทนที่จะเป็นลำดับของคำเดี่ยว ด้วยแนวคิดดังกล่าวจึงได้มีการพัฒนาเทคโนโลยีใหม่ในการค้นคืน ซึ่งมีชื่อว่า "Essence Extractor Engine" หรือย่อว่า SENTRAX [SENT04] ดัชนีของการค้นคืนถูกสร้างจากการสร้างความสัมพันธ์ของคู่ของคำที่มีความหมายและอนุญาตให้สร้างนิยามและส่งผ่านแนวคิดในการค้นหาจากคำที่ถูกเลือก ซึ่งที่คำเหล่านั้นถูกเลือกโดยอิสระต่อกัน เทคโนโลยีนี้ถูกใช้ในวิทยานิพนธ์นี้ ในการค้นคืนข้ามภาษา จะส่งผ่านของคอนเซ็ปแทนที่จะแปลคำที่ใช้ค้นหาทีละคำสามารถลดโอกาสในการเกิดความกำกวมของการแปล คอนเซ็ปจะถูกรักษาไว้โดยความสัมพันธ์ร่วมกันของการแปลคำต่างๆในคอนเซ็ปเองหรือกับคำที่ประกอบรอบข้างในตัวเอกสาร ผลและผลสืบเนื่องได้ถูกวิเคราะห์และแสดงเอาไว้ในวิทยานิพนธ์ ฟังก์ชั่นอื่นๆของ SENTRAX-Engine (เช่น การยืดหยุ่นในการค้าหาคำ การค้นหาโดยเปรียบเทียบความคล้ายคลึงกันของเอกสาร) รวมถึงการติดต่อระหว่างผู้ใช้กับระบบผ่านกราฟฟิค (Graphic User Interface) ได้พิสูจน์ตัวเองถึงความได้เปรียบในการใช้งานที่มีอยู่

กระบวนการขั้นต้นที่จำเป็นและเทคนิคเทคนิคถูกออกแบบเพื่อใช้เชื่อมโยง ๒ กลุ่มดัชนีสำหรับการค้นคืนข้ามภาษาเข้าด้วยกัน ส่วนสำคัญ ๓ ส่วนได้แก่ ๑. กระบวนการขั้นต้น ซึ่งจะรับผิดชอบในการปรับโครงสร้างของคำในต่างภาษาให้เทียบเคียงกันได้ ๒. สะพานที่จะเป็นทางผ่านของคำที่ใช้ในการค้นหาจากภาษาต้นทางไปยังภาษาเป้าหมาย ๓. เครื่องมือวัดความคล้ายคลึงกันของคอนเซ็ป ซึ่งจะควบคุมความสมดุลย์ของคอนเซ็ปหลังจากการแปล ณ. เวลานี้ ทั้งสามส่วนยังไม่สามารถทำงานได้อย่างอัตโนมัติ แต่สามารถทำงานได้แบบแมนวล ถึงแม้ว่าระบบจะยังไม่สมบูรณ์ เราก็สามารถทดสอบสมมุติฐานต่างๆได้

จากผลของของการทำการทดสอบเราสามารถมั่นใจได้ว่า การค้นคืนข้ามภาษาด้วยคอนเซ็ปเน็ตเวิร์ทสามารถให้ผลที่ดีได้ ความกำกวมของการแปลสามารถถูกทำให้ลดลงได้โดยการพิจารณาความสัมพันธ์ร่วมกัน การเรียนรู้ในระหว่างการค้นคืนจะช่วยให้ผู้ใช้ตัดสินใจได้ว่า การค้นหานั้นมีความเป็นไปได้ที่จะได้ข้อมูลที่ต้องการมากน้อยเพียงใด เพราะบางทีข้อมูลดังกล่าวอาจจะไม่ได้อยู่ในฐานข้อมูลก็ได้ คำต่างๆที่สอดคล้องกับคำที่ใช้ค้นหาถูกนำเสนอผ่านกราฟฟิคเพื่อให้ผู้ใช้ได้เลือกนั้นมีบางส่วนที่สามารถเทียบเคียงกันได้ ด้วยปรากฏการณ์นี้ ทำให้เราสามารถยืนยันได้ว่า เอกสารที่เชื่อมโยงกับคำดังกล่าวมีความเป็นไปได้สูงที่จะกล่าวถึง เรื่องราวเดียวกันหรือใกล้เคียงกัน

INHALTVERZEICHNIS

1 ÜBERBLICK

Der Bedarf an Informationen und deren Austausch steigt Tag für Tag. Das trifft nicht nur für die einheimische Sprache, sondern auch für Informationen in fremder Sprache zu. Mit der Zunahme des Datenumfangs erhöht sich auch der Bedarf an geeigneter Suchtechnologie. Bei einer herkömmlichen monolingualen Suche wird die benötigte Information durch die vom Nutzer formulierte Suchanfrage (innerhalb eines Information-Retrieval Modells) erreicht. In der vorliegenden Untersuchung wird ein duales IR-Modell präsentiert (siehe Abschnitt 2.2). Anhand von Hilfstechniken (siehe Abschnitt 2.4), z.B. Relevanz-Feedback, kann die Suchanfrage erweitert werden, indem Zusatzwörter zu den Suchwörtern anteilig hinzukommen. Die große Herausforderung in der monolingualer Suche ist die Verbesserung der IR-Methode, meist in einem konventionellen Modell, um den besten Precision- bzw. Recallwert zu erreichen. Auf der anderen Seite versuchen die Forscher die Suchverhältnisse besser zu verstehen, damit der Bedarf des allgemeinen Suchers durch die Mensch-Maschine-Schnittstelle bzw. Suchmethode am besten erfüllt werden kann. Ein neuartiger Ansatz hier wird durch die Suchanwendung „SENTRAX“ (Kurzform für "Essence Extractor Engine") gegeben, die dem Nutzer eine Interaktion während der Suche anbietet. Besondere Vorteile der SENTRAX sind durch die Lernmöglichkeit während des Suchprozesses und die Ideen- bzw. Begrifferweiterung mittels einem Konzeptnetzes (siehe Abschnitt 2.5.1.3 und 2.5.2.2) gegeben, was bei der konventionellen Suchmethode nicht vorliegt. Weil der Sucher mit der SENTRAX auf der Konzeptschicht statt der Wortschicht arbeitet, ist es einfacher, weiterführende Zusatzbegriffe auszuwählen, und zwar nicht aus der Luft gegriffene, um so das Suchkonzept zu verschärfen und dabei nicht von der Suchrichtung abzulenken.

Das klassische Vorgehen, eine bilinguale Suche aufzubauen, ist die Verknüpfung von zwei IR-Systemen durch eine Übertragungsmethode (siehe Abschnitt 3.1.2). Das größte Problem dabei ist die Mehrdeutigkeit der Übersetzung. Die Mehrdeutigkeit kommt aus den vielfältigen verschiedenen Übersetzungsmöglichkeiten. Es ist beim Computer schwer zu entscheiden, welche Übersetzung am besten zum gegebenen Kontext passt. Ein weiteres Hindernis bei manchen Sprachen liegt im Mangel an Ressourcen für die

Entwicklung, z.B. der Mangel an elektronisch lesbarem Wörterbuch für die Sprache. Obwohl die Umsetzungssprache (siehe Abschnitt 3.1.2.6) solche Mängel umgehen kann, erhöht sich bei wiederholten Übersetzungen trotzdem die Chance der Mehrdeutigkeit.

Man benutzt ein Konzept oft, um etwas zu definieren, damit die beabsichtigte eindeutige Bedeutung erkannt und verstanden werden kann. So ein Konzept soll aus mehreren Eigenschaften zusammengesetzt sein. Dadurch wird es möglich, dass das gesamte Konzept noch begriffen wird, obwohl einige Eigenschaften in der Beschreibung fehlen. Anhand dieser Vorgaben könnte mittels einer toleranten Konzeptübertragung die Mehrdeutigkeiten bei der bilingualen Suche vermieden werden, weil sich der Kern des Konzeptes nach der Übertragung noch erhält.

Stellen Sie sich vor, Sie wären in ein Land gereist, wo Sie kein Wort verstehen. Sie suchten dringend ein WC und hätten jemanden mit der Körpersprache gefragt. Die Antwort wäre natürlich auch in der Körpersprache, aber Sie könnten das verstehen. Die Handlungen, die Sie gemacht bzw. gesehen hätten, sind die Eigenschaften der Frage bzw. Antwort. Diese müssen vielleicht nicht vollständig klar sein, können aber trotzdem auf das bestimmte, gewünschte Konzept hinweisen. Dies wäre ein Beispiel für eine Konzeptübertragung.

In unserem Fall von Textkorpora als Träger der Begriffe und Informationen kann das Modell und das Vorgehen schematisch wie in Abbildung 1 dargestellt werden. Die einzelnen Probleme, Schritte, Maßnahmen und Ergebnisse werden im Detail in Kapitel 4 und 5 dargelegt.

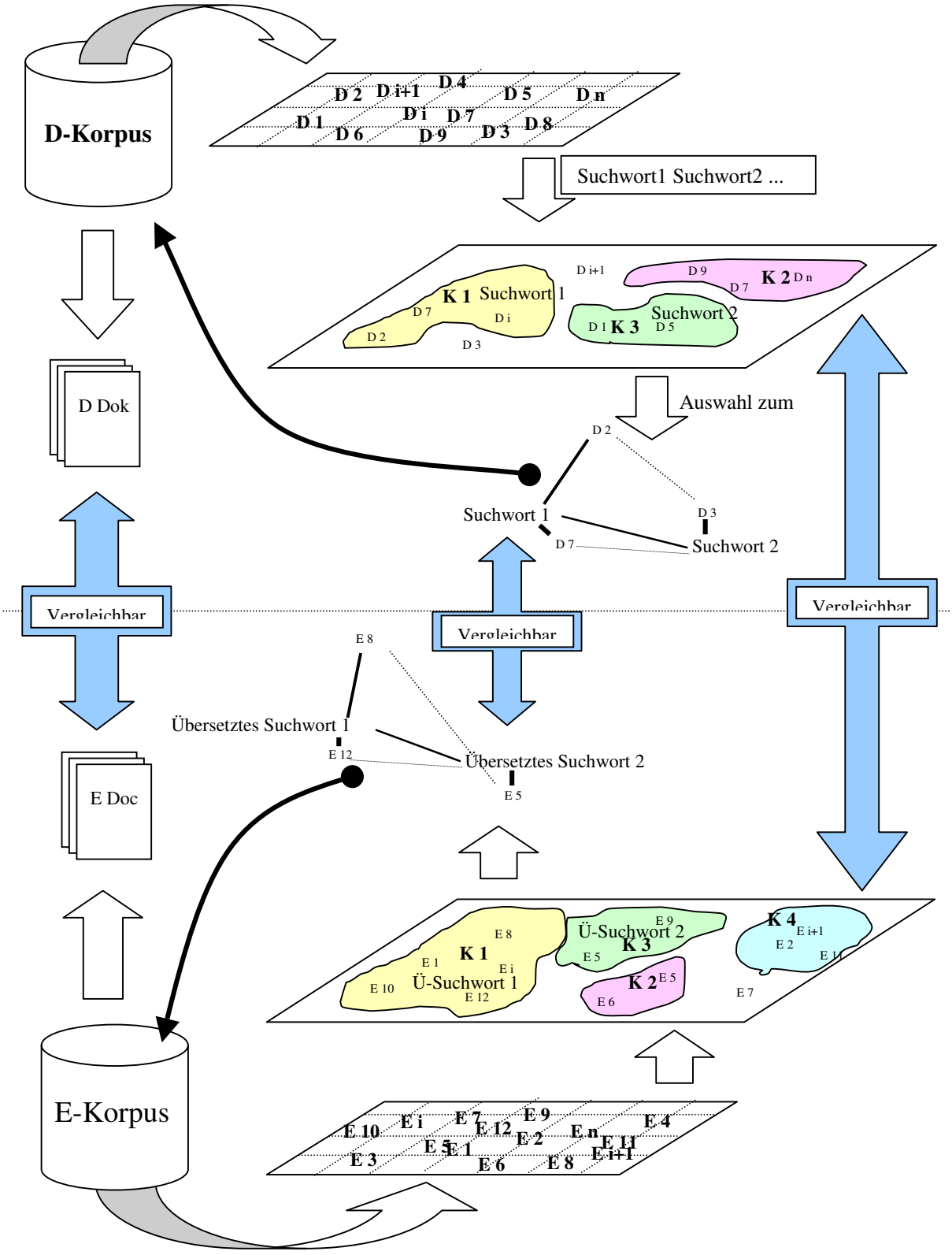


Abbildung 1 Grundidee der bilingualen Suche mittels Konzeptnetz

Bislang steht für die deutsche monolinguale Suchanwendung, die Konzeptnetze verwendet, die SENTRAX zur Verfügung. Hier wird zum anfänglichen Konzept des Nutzers ein passendes im SENTRAX-Container gefunden und im Verlauf der Interaktion wechselseitig angenähert. Die überflüssigen, unwichtigen Wörter auf der grafischen Darstellung können ein Hindernis bei der Auswahl sein. Für die Beseitigung der wenig Information tragenden Worte kann man eine Tagger-Anwendung einsetzen. Dies ermöglicht verschieden starke Eingriffe in die zu indexierenden Texte. Beispielsweise kann eine Stammformreduzierung besonders interessant sein für eine Sprache, die viele abgeleitete Formen hat. So lassen sich auch Grundformen der unterschiedlichen Sprachen angleichen. Für derlei Verarbeitungen stand hier das Programm TreeTagger[1] zur Verfügung. Wortarterkennung und Stammformreduzierung wurde zwar genutzt, aber nur mit manueller Hilfe, da die Module aus technischen Gründen für die vorliegende Untersuchung noch nicht vollständig in die Anwendung integriert wurden. Aus den Experimenten lässt sich vermuten, dass die Stammformreduzierung eine merkbare Auswirkung auf die in Frage kommenden Assoziationen hat (siehe Abschnitt 4.1.1).

Um die bilinguale Suche mit der SENTRAX aufzubauen, werden zwei getrennte Container unabhängig voneinander erstellt. Insofern sind beide Richtungen zunächst gleichwertig. Bei der jeweils begonnenen Suche können die in den Dokumentmengen vorhandenen Zusammenhänge durch den direkten und indirekten Assoziationsprozess betrachtet werden. Eventuelle Tippfehler bzw. Schreibvarianten werden durch die Lexico-Funktionen der SENTRAX aufgedeckt. Weitere Funktionen erlauben eine grafische Darstellung von Wortmengen. Auf die Parameter der Darstellungen (Wortanzahl, Dokumentenanzahl etc.) kann während der Suche zugegriffen werden. Um beim bilingualen Suchschritt die Begriffe der Ausgangsprache in die Zielsprache zu übertragen, stehen zwei Möglichkeiten zur Verfügung, und zwar das elektronisch lesbare deutsch-

[1] IMS, Universität Stuttgart - http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

englische Wörterbuch von TU-Chemnitz[2] und die Transfermatrix, die der Methode von Rapp [RAPP99] angelehnt ist (siehe Abschnitt 4.2.3). Man wird in dieser Arbeit sehen, dass die Übertragung der korpusbasierten Begriffe das gesamte (jeweilige) Suchkonzept bewahrt. Dann werden die Konzeptvergleichsmaße entworfen. Die im Hintergrund stehende Theorie für konzeptionelle Graphabgleichung ist hier an die Vorgaben von [MLG00] angelehnt. Diese Vergleichsmaße dienen zur besseren Auswahl von Suchkonzepten.

Die Bausteine des bilingualen Systems werden mit der SENTRAX passend zusammengestellt. Allerdings sind noch nicht sämtlich ablaufenden Schritte automatisiert, der Benutzer muss hier und da eingreifen. Das ist im gegenwärtigen Stadium der Entwicklung auch verständlich, da zunächst die Brauchbarkeit bzw. Korrektheit der Arbeitshypothesen geprüft und gesichert werden soll und erst danach programmtechnische Festlegung (von Parametern, linguistischen Tools, Programmabläufen, Übergabe von Zwischenergebnissen usw.) für ein Komplettsystem erfolgen.

Die Hypothesen wie die bilinguale Suche mittels Konzeptnetzen unter verschiedenen Bedingungen mit einem parallelen Korpus reagiert, lassen sich hier aber mit dem vorhandenen System überprüfen. Auch kann das Änderungsverhalten des Konzeptnetzes bei Containervergrößerung beobachtet werden. Das Ergebnis ist, dass das Konzeptnetz ziemlich stabil aussieht, wenn der Container groß genug ist und das entsprechende Thema in den relevanten Dokumenten nicht wechselt. Ein nicht-paralleler Korpus wird dann aus dem parallelen Korpus aufgebaut, um die normale Situation der bilingualen Suche testen zu können. Die Experimente beim nicht-parallelen Korpus ergeben, dass die Suche mittels Konzeptnetz eine leistungsfähige Methode für das bilinguale bzw. multilinguale Informationsretrieval ist.

Schlussfolgerungen aus den Untersuchungen und ein Ausblick finden sich in Kapitel 6. Die durchgeführten Arbeiten mit der bilingualen Suchanwendung der vorliegenden

[2] unterladen von ftp://ftp.tu-chemnitz.de/pub/Local/urz/ding/de-en/ Stand Nov. 2005

Form geben auch Hinweise für einig künftige Verbesserungen, die dann in einem Komplettsystem verankert werden können.

2 MONOLINGUALE SUCHE

2.1 Einführung in das Informationsretrieval

Der Begriff Informationsretrieval (IR) setzt sich aus zwei Grundwörtern zusammen, und zwar „Information“ und „Retrieval“. Mit „Information“ ist hier mehr gemeint als das, was man mit dem Wort „Datei“ bezeichnet, weil die Information (lat. informare = bilden, durch Unterweisung Gestalt geben) potenziell oder aktuell vorhandenes, nutzbares oder genutztes Wissen sei, während eine Datei nur Aufbewahrungsort des geschriebenen Textes ist. Das Retrieval stellt den Abruf der Information entsprechend der Anfrage aus der Dokumentensammlung bzw. aus der Datenbank dar. Deshalb werden zwei zusammenhängende Hauptaspekte bedeutsam, nämlich die (gegebene) Dokumentsammlung und die (aktuelle) Anfrage vom Nutzer. Um das Informationsretrieval zu verwirklichen, benötigt man das IR-Modell (siehe unten). Die Informations-, die Anfragenorganisation und die Relation zwischen Informationen und Anfragen werden durch das Modell veranschaulicht. Ein anderer wichtiger Teil ist die Assoziation zwischen bestimmten Anfragen und den entsprechenden Dokumenten. Hierfür ist die Einordnungsfunktion verantwortlich. Dieser hauptsächliche Aspekt wird umfangreich verwendet.

Zwei Richtungen der Informationsretrieval-Forschung lassen sich ausmachen, und zwar

- Systementwicklung – es geht darum, dass die Forscher mit ihrem IR-System die beste Leistung erzielen. Die Güte wird durch den Recall und die Precision gemessen.

- Studium des Nutzerverhaltens – es geht darum, dass die Forscher sich dafür interessieren, wie der Nutzer seine Anfrage formuliert oder wie er mit einem System umgeht. Mit dem Verständnis dieses Geschehens könnte man System und Nutzer besser verbinden und einem Nutzer die erforderlichen Informationen leichter bereitstellen.

Nach [BEAZ99] wird das Quadrupel des IR-Modells als $\lfloor D, Q, F, R(q_i, d_j) \rfloor$ definiert, wobei D die Menge der Dokumente in der Sammlung aus der logischen Sicht, Q die Menge der benötigen Information des Nutzers aus der logischen Sicht, F das Bezugssystem für die Um-

formung der Dokumente, der Anfrage und ihrer Relation und $R(q_i, d_j)$ die Ranglistenfunktion ist, die die Assoziation der Anfrage $q_i \in Q$ und $d_j \in D$ repräsentiert.

Die Art von Dokumenten können wir in zwei Gruppen unterteilen, und zwar statische Dokumente, die praktisch Archivcharakter haben und dynamische Dokumente, die öfter in der Sammlung erneuert werden oder dazukommen, beispielweise jede Woche, jeden Tag oder jede Stunde. Die IR-Methoden, die hier wirken, sind ad hoc (Konzept für statische Dokumente mit dynamischer Anfrage) und Filtrieren (Konzept für statische Anfrage mit dynamischen Dokumenten). Der Unterschied zwischen ad hoc und Filtrieren besteht darin, dass in der ad hoc Gruppe die Dokumente entsprechend der neuen Anfrage aus der Sammlung ausgegeben werden, während beim Filtrieren die neuen Dokumente durch das Nutzerprofil geleitet werden, um die brauchbaren Dokumente herauszufiltern. Das Nutzerprofil besteht meist aus einfachen Stichwörtern.

In diesem Kapitel werden einige klassische IR-Modelle kurz vorgestellt. Auch werden Leistungsvermögen und Hilfstechniken beschrieben. Der wichtigste Teil ist die SENTRAX-Anwendung, die in unseren Ansatz verwendet wird.

2.2 Modelle

Viele Modelle sind schon lange bekannt und bereits "klassisch". Von Zeit zu Zeit gibt es auch neue Vorschläge, doch sind manche von den alten noch funktionstüchtig und praktisch im Einsatz.

2.2.1 Boolesches Modell

Dieses Modell basiert auf der Mengentheorie. Die Relation zwischen Indexen und Dokumenten ist durch ein binäres Gewicht definiert: $w_{i,j} = \{0,1\}$. Die Anfrage wird mit Hilfe der zugrunde liegenden Mengen operatoren – „und“, „oder“, „nicht“ – formuliert. Die Anfragen werden in die disjunktiven Normalform (Abk. DNF) überführt. Bei dieser Form wird ein bi-

närer Indexierungsvektor betrachtet, wobei „0“ bedeutet, dass die Anfrage auf den Index nicht zugreift, und „1“, dass sie dieses macht.

Das Boolesche Modell beurteilt das Dokument entweder als relevant oder als irrelevant. Aufgrund dieses Urteils wird ein Dokument, das der Anfrage nur zum Teil entspricht, in die Gruppe irrelevanter Dokumente eingeordnet. Der Vorteil dieses Modells besteht in der einfachen und klaren Form.

2.2.2 Vektormodell

Anstatt des binären Gewichtes erlaubt das Vektormodell, ein nichtbinäres Gewicht zwischen den Schlagwörtern bzw. Indextermen und dem Dokument zu nutzen. Das bekannteste Gewichtschema ist das „tf-idf“ Schema, das über den tf-Faktor und den idf-Faktor berechnet wird (siehe Formel 1; Anhang 7.5).

Das Ähnlichkeitsmaß identifiziert die Stärke der Relation zwischen der Anfrage und dem Dokument. Das bekannteste Vergleichsmaß der Ähnlichkeit ist der Kosinuswinkel bzw. das Skalarprodukt (siehe Formel 4; Anhang 7.5).

Zusätzlich wurde ein Gewicht zwischen der Anfrage und der Indexe eingeführt und empfohlen.

Es ist deutlich, dass das Vektormodell manchmal vorteilhafter ist, weil auch nur zum Teil passenden Dokumente entsprechend der Anfrage aufgerufen werden. Außerdem werden die Dokumente durch den Kosinuswinkel geordnet. Der Nachteil besteht darin, dass das Vektormodell für abhängige Indexierungen nicht verwendet werden kann, sondern nur bei gegenseitig unabhängigen Indexierungen.

2.2.3 Wahrscheinlichkeitsmodell

Unter dem Wahrscheinlichkeitsmodell wird das IR-System im Wahrscheinlichkeitsrahmen betrachtet. Die Grundidee besteht darin, dass die Anfrage des Nutzers einer Menge relevanter

Dokumente entspricht. Die Menge der relevanten Dokumente wird ideale Antwortmenge genannt. Natürlich gibt es einzelne Eigenschaften der idealen Antwortmenge abhängig von der Anfrage. In der realen Situation sind die Eigenschaften zur Laufzeit der Anfrage nicht bekannt. Dadurch kann man die volle ideale Antwortmenge derzeit nicht finden. Der Ähnlichkeitskoeffizient bzw. das Ähnlichkeitsmaß zwischen Anfragen und Dokumenten wird beim Wahrscheinlichkeitsmodell als Wahrscheinlichkeit angegeben, in der sich die Relevanz zwischen Dokument und Anfrage ausdrückt.

Zuerst wird der anfängliche Wahrscheinlichkeitswert der Beziehung zwischen Index k_i und Dokumentsmenge *R* angegeben, damit die erste Dokumentenmenge abgerufen werden kann. Die Relevanz wird durch die Wahrscheinlichkeit abgeschätzt, dass ein Begriff in einem gegebenen Dokument vorkommen soll, statt nur durch die Anwesenheit oder die Abwesenheit des Begriffs in einem Dokument. Anhand der Beeinflussung des Wahrscheinlichkeitswertes durch die Eingaben des Nutzers entwickelt sich der Wahrscheinlichkeitswert der Beziehung. Dieser Prozess lässt sich wiederholen, damit der Wahrscheinlichkeitswert der Beziehung verbessert wird. Durch den Wiederholungsprozess sollte sich der Wahrscheinlichkeitswert der wirklichen Beziehung zur idealen Antwort annähern.

Für das Wahrscheinlichkeitsmodell wird das Gewicht des Indexes binär gesetzt, wobei es $w_{i,j} = \{0,1\}$ und $w_{i,q} = \{0,1\}$ für das Dokument j und die Anfrage q in die Untermenge des Indexes k sind. Gegeben seien R als Menge der relevanten Dokumente und $\overline{R}$ als Menge der irrelevanten Dokumente. Damit wird der anfängliche Wert definiert (siehe Formel 6; Anhang 7.5).

Um die Rangliste der erhaltenen Dokumente zu bekommen, wird die Ähnlichkeit zwischen dem Dokument und der Anfrage berechnet (siehe Formel 9; Anhang 7.5). Die Dokumente in der Rangliste werden nach dem Ähnlichkeitswert sortiert.

Der wesentliche Vorteil dieses Modells ist die Sortierung der Dokumente in der Rangliste nach ihrer Relevanzwahrscheinlichkeit. Aber es gibt auch Nachteile, nämlich: dass (1) die anfänglichen relevanten und unrelevanten Dokumentmengen zunächst auf ihre Größe hin geschätzt werden müssen, (2) die Indexe auf der angenommenen Unabhängigkeit basieren müs-

sen, (3) die Häufigkeit der Relationen zwischen den Indexausdrücken bzw. den Begriffen und dem Dokument in dem Prozess nicht verwendet wird. Die grundlegenden Probleme des Wahrscheinlichkeitsmodells sind die Parameterabschätzung, bei der die guten Dateien zu trainieren sind, und die Nutzung der Unabhängigkeitsvermutung, bei der der Zusammenhang zwischen einigen Wörter getrennt wird. In Online-Systemen wird die Relevanz-Feedback-Technik verwendet, um den anfänglichen Wert zu initialisieren. Das Problem der Unabhängigkeit führt zum deduktiven Netzwerkmodell und der logistischen Regression [GRFR98]. Andere Varianten des Wahrscheinlichkeitsmodells kamen dazu, z.B. unbinäres Unabhängigkeitsmodell, Poissonmodell, Begriffskomponentenmodell.

2.2.4 Erweitertes boolesches Modell

Bei diesem Modell werden die Ideen vom booleschen Modell und Vektormodell zusammengeführt. Wegen der Vorteile des booleschen Modells, das einfach und klar ist, und des Vektormodells, das schnell und leistungsfähig ist, werden diese zusammengebracht. Beim erweiterten booleschen Modell werden die Dokumente in den t-dimensionalen Hyperkubus gelegt, wobei t die Anzahl aller Indexe ist. Jede Achse repräsentiert das jeweilige Stichwort bzw. den jeweiligen Index. Die Position des Dokumentes hängt von seinem Gewicht bezüglich des Stichwortes ab.

Das erweiterte boolesche Modell ist ein Hybridmodell, das die Eigenschaft der booleschen Operation und die Eigenschaft der algebraischen Distanz einschließt. Obwohl dieses Modell schon vor über 20 Jahren angekündigt wurde, wird es nicht oft verwendet.

2.2.5 Verallgemeinertes Vektorraummodell

Wenn die Indexausdrücke in der Dokumentsammlung t verschiedene Werte haben, $k_1, k_2, \ldots, k_t$, und die Gewichte zwischen den Indexausdrücken und den Dokumenten definiert sind, bezeichnet man das Paar aus Index und Dokument als (k_i, d_j). Wenn das Gewicht binär ist, sind die Bitmuster des Gewichts 2^t minimale Ausdrücke, $m_1 = (0,0,\ldots,0)$,

$m_2 = (1,0,\ldots,0)$, $m_{2^t} = (1,1,\ldots,1)$. Jedes Dokument entspricht einem Muster von 2^t minimalen Ausdrücken. Die Zahl „1“ auf dem Platz i repräsentiert die Indexierungsausdrücke k_i.

Da das Vektormodell auf der Annahme der Unabhängigkeit basiert, versucht man die Unabhängigkeit der Indexe nachzubilden, d.h. für zwei beliebige Indexvektoren k_i und k_j muss ihr Skalarprodukt gleich null sein.

Für die Berechnung der Ähnlichkeit kann man den Standardkosinus verwenden. Obwohl dieses Modell eine interessante Idee ist, ist die Berechnungszeit auf einer großen Dokumentsammlung sehr groß.

2.2.6 Latent Semantik-Indexierungsmodell

Bei diesem Modell stehen die Bemühungen im Vordergrund, die Anfrage und die Dokumente in den Semantikraum, sogenannter Konzeptraum, zu übertragen, statt die in der Anfrage gegebenen Wörter zu nehmen. Weil das gleiche Konzept von mehreren verschiedenen Wörtern dargestellt werden kann, kann es vorkommen, dass die Wörter in der Anfrage nicht unbedingt im Dokument auftauchen. Bei diesem Modell interessiert man sich nur dafür, dass die Anfrage in der gleichen Umgebung im Semantikraum landet wie die Dokumente. Wenn das passiert, kann man die entsprechenden Dokumente, die vom gleichen Thema wie die Anfrage handeln, erhalten.

Das zugrunde liegende Prinzip des latenten semantischen Indexierungsmodells (Abk. LSI-Modell) ist die Tatsache, dass die Dokumente und Anfragevektoren auf einem kleindimensionalen Vektorraum abgelegt werden.

Gegeben sei $\mathbf{M}_{t\times N} = (m_{ij})$, die assoziierte Matrix mit den *t* Indexausdrücken und *N* Dokumenten. Der Wert von m_{ij} ist das Gewicht w_{ij}, das durch die tf-idf-Gewichtstechnik wie beim Vektorraummodell berechnet wird.

Beim LSI-Modell wird der Vektor $\mathbf{M}$ durch die Singularwertzerlegung in drei Komponenten zerlegt (siehe Formel 14; Anhang 7.5).

Die Zahl *s*, *s*<*r*, wird für den reduzierten Semantikraum gewählt, damit die Anzahl der Dimensionen des Konzeptes reduziert werden kann. Bei der Auswahl muss das Gleichgewicht der wechselseitigen Effekte erhalten bleiben:

1. *s* muss groß genug sein, damit alle im Sinne der mathematischen Vorstellung semantischen Auffassungen in einer realen Datei repräsentiert werden können.
2. *s* muss klein genug sein, damit alle unrelevanten repräsentierenden Details herausgefiltert werden können.

Den optimalen Wert *s* in der Formel 15 (siehe Anhang 7.5) bekommt man nach den ausgeführten Experimenten. Das Ergebnis von der Reduzierung der Singularwertmatrix ist die Matrix $\mathbf{M}_s$ mit dem Rang *s*, die beinahe die originale Matrix im Sinn der „least square“ Methode ist.

Wenn das System die Anfrage erhält, wird das Pseudo-Dokument durch die originale Matrix herausgesucht. Wenn das Pseudo-Dokument aus der Spalte k käme, werden die Dokumente aus der Korrelationsmatrix $\mathbf{M}_s^t\mathbf{M}_s$ aus Zeile k herausgenommen. Die Dokumente in der Zeile k werden in der Rangliste nach dem Wert sortiert.

Obwohl das latente semantische Indexierungsmodell einen Vorteil gegenüber der Nutzung der Rechenmatrix hat, und das Ergebnis bei verschiedenen Kollektionen in vielen Berichten besser als die konventionellen Modelle ist, entsteht auch ein Nachteil, und zwar die Laufzeitverlängerung. Im Vergleich zum Vektormodell oder Wahrscheinlichkeitsmodell mit invertiertem Index benötigt das latente semantische Indexierungsmodell mehr Rechenzeit. Für die Sammlung mit *N* Dokumente und die Singularwertzerlegungsmatrix $\mathbf{S}_s$ mit Rang *s* wird die Rechenzeit durch $O(N^2s^3)$ abgeschätzt.

2.3 Vergleichsmaße

2.3.1 Recall und Precision

Für eine Anfrage q_i gibt es eine bestimmte Menge an relevanten Dokumenten in der angeforderten Antwort, die wird $Relevant(q_i)$ genannt. Bei unterschiedlichen Retrievalsystemen wird die Retrievalmenge der gleichen Anfrage in verschiedenen Listen eingetragen. Gegeben sei die Retrievalmenge $S_{Ret}(q_i)$ der Anfrage q_i vom Retrievalsystem *S*. Je mehr Dokumente in der angeforderten Antwort $Relevant(q_i)$ in der Retrievalmenge $S_{Ret}(q_i)$ von dem Retrievalsystem S gefunden werden, desto besser ist das Retrievalsystem.

Die zu messende Leistung eines Retrievalsystems ist vom Recall und von der Precision abhängig.

$$\text{Recall} = \frac{|S_{Ret}(q_i) \cap Relevant(q_i)|}{|Relevant(q_i)|}$$

$$\text{Precision} = \frac{|S_{Ret}(q_i) \cap Relevant(q_i)|}{|S_{Ret}(q_i)|}$$

wobei $|\bullet|$ die Funktion ist, die die Anzahl der Elemente der Menge ausgibt.

Weil die Anzahl der Anfragen in einer Batch-Anfrage nicht immer eins und die Anzahl der Dokumente in einer Antwortmenge entsprechend einer Anfrage q_i nicht unbedingt gleich ist, wird der Standard für Recall-Precision definiert. Um den Standard zu bilden, werden die elf Punkte des Recallwertes zuerst eingestellt, d.h. 0%, 10%, 20%, ..., 100% als Recallwert. An jedem Punkt wird die Precision berechnet. Für die berechneten Precisionwerte von allen Anfragen q_i wird noch einmal der Durchschnitt berechnet (siehe Formel 16; Anhang 7.5).

Weil die Anzahl der Dokumente von der Antwortmenge $Relevant(q_i)$ entsprechend einer Anfrage q_i nicht gleich ist, wird der Precision auf jeder Recallstufe $recall_j$, j = 0, 1, 2,...,10 für

die Recallstufe 0%, 10%, 20%,...,100%, wie durch die Formel 17 (siehe Anhang 7.5)berechnet.

Diese Precision-Recall-Bewertung ist sehr populär geworden. Zwei der vier in [BEAZ99] genannten Nachteile sind, dass die erforderliche zuweisende Kenntnis von allen Dokumenten in einer Sammlung zum maximalen Recall-Wert führt und das Funktionieren des Recall-Precision-Vergleichsmaßes nur im Batch-Verfahren gilt.

2.3.2 Weitere Parameter und Methoden

Die Bewertung des Singulärwertes ist das Leistungsvergleichsmaß des Retrievalalgorithmus, in dem der Einfluss von jeder Anfrage bewertet wird. Dadurch erfährt man, welche Anfrage das Ergebnis positiv oder negativ beeinflusst hat und passt den Retrievalalgorithmus an. Weitere wichtige Parameter und Methoden stellen die R-Precision, das Precision- Histogramm, das E-Maß dar (vgl. [BEAZ99]).

2.4 Hilfstechniken

Das IR-Modell spielt in einem Retrievalsystem deswegen eine große Rolle, weil es als Formulierungsstrategie zwischen der Dokumentsammlung und der Anforderung des Nutzers angewandt wird. Es gibt aber zusätzliche Strategien, die man mitnutzen kann, ohne Auswirkungen auf das IR-Modell zu haben, mit dem IR-Modell zu geben, sogenannte die IR-Hilfstechnik. Die IR-Hilfstechniken können in zwei Anwendungsarten unterschieden werden, nämlich in Anfrageverfahren und Dokumentverfahren.

Die Beispiele von der IR-Hilfstechnik sind

- Relevanz-Feedback – die Dokumente in der Spitze *k* der Rangliste oder in der Auswahl des Nutzers werden bei herkömmlicher Methode als relevante Dokumente bezeichnet. Die Ausdrücke in den vermutlichen relevanten Dokumenten werden herausgenommen, um sie in der Anfrage hinzuzufügen. Die neue Anfrage, die alte mit zusätzlichen Ausdrücken, wird noch mal in den Suchprozess hereingebracht.

- Aufteilung der Dokumente oder der Ausdrücke – die Dokumente oder die Ausdrücke werden in die Gruppe entweder automatisch oder manuell geteilt. Das Ziel ist die unrelevanten Dokumente früh auszuscheiden. Nur die ziemlich relevanten Gruppen werden gefiltert.
- Passage-basiertes-Retrieval – in den relevanten Dokumenten gibt es oft einen unrelevanten Teil. Der relevante Teil bzw. die Passage ist irgendwo ziemlich verdichtet. Insofern wird die Anfrage direkt der Passage angepasst. Die Ergebnisse von allen Passagen im Dokument werden zu einem eigenen Ähnlichkeitskoeffizienten vereinigt.
- Syntaxanalyse – dies gehört zur Analyse der Nomengruppe, der Stammform, usw. Die Berechnung zur Identifikation eines Begriffs und zusammengehöriger Satzteile ist nützlicher als die Analyse der Wörter in der Nachbarschaft. Die Syntaxanalyseregeln oder Listen der Ausdrücke werden verwendet, um die korrekten Ausdrücke wie „New Delhi“ zu erkennen. Außerdem kann die Stammformreduzierung das IR-System unterstützen, um die zahlreichen abgeleiteten Formen auszugleichen.
- N-Grams – die Anfrage wird in dem Fenster der Größe N betrachtet, das die Reihenfolge von Zeichen abdeckt. In OCR-Systemen werden N-Grams verwendet, um die Fehler von der Mustererkennung oder Schreibung zu korrigieren. Diese Strategie ist von der Sprache abhängig.
- Thesauri – die Thesauri können von den Texten oder durch eine manuelle Methode erzeugt werden. Sie erweitern entweder die Anfrage oder die Dokumente, um das gesamte Retrieval zu verbessern.
- Semantisches Netz – das semantische Netz ist das Konzeptnetz, das sich in einer hierarchischen Form bindet. Die Verbindung zwischen den Begriffen wird von der Stärke einer Beziehung bestimmt. Das semantische Netz wird zur Erweiterung der Anfrage oder des Dokumentes angewendet, indem dazu andere verwandte Begriffe betrachtet werden.

2.4.1 Anfrageverfahren

Bei der Kommunikation gibt es drei wichtige Teile, nämlich Sender, Empfänger und Nachrichten. Eine gute Leistung der Kommunikation wird erbracht, wenn der Sender die optimale Nachricht formuliert, die der Empfänger verlangt und einfach verstehen kann. Die Frage von dem Empfänger an den Sender gehört zur Kommunikation dazu, denn es ist wichtig, wie er ihn fragt und ob der Sender die Frage richtig versteht. Bei der Informationssuche ist es ähnlich, denn man vergleicht ein Suchsystem mit einem Sender, das von einem Retrievalsmodell unterstützt wird und die Bedeutung der Frage nicht verstehen kann. Es ist bei der Informationssuche schwieriger als bei der normalen Kommunikation, weil die Fähigkeit des Senders beschränkt wird. Wenn der Nutzer eine Retrievalmethode gut versteht, kann er geeignete Anfrage formulieren und die richtigen Dateien aus der Sammlung einfach bekommen. Wenn er aber nicht daran gewöhnt ist, gäbe er wahrscheinlich verworrene Anfragen und bekäme verwirrende Dateien.

Die meisten Nutzer haben keine Ahnung von der Retrievalmethode und kennen die Dokumentensammlung nicht so gut oder vielleicht gar nicht. Die Anfrage kann als wichtiger Schlüssel zur Lösung betrachtet werden. Die optimale Anfrage könnte die erforderlichen Dokumente zurückliefern. Einige Nutzer verbringen viel Zeit, um eine gute Anfrage zu formulieren. Die Retrievalzeit ist nicht nur der Retrievalprozess, sondern auch die Zeit, um eine gute Anfrage zu finden. Die meisten Nutzer verlieren viel Zeit bei diesem Schritt.

Grootjen und van der Weide haben in [GRWE02] auf die Schwierigkeit der Anfrageformulierung abgehoben. Sie haben zwei Fragen gestellt, und zwar, ob man richtig weiß, was man gerade sucht und wenn man es weiß, wie man eine optimale Anfrage formulieren kann, um seine gewünschte Information zu erhalten. Eine gute Anfrage erfordert, dass der Nutzer irgendwie vorhersehen kann, welche Ausdrücke in den benötigten Dokumenten stehen. Das heißt, er muss die Dokumentensammlung gut kennen.

Die automatische Formulierung der Anfrage durch die Expansion der Anfrage oder andere Methoden hilft dem Nutzer, um die benötigen Dateien leichter zu finden. Es erhöht die Chance, bisher nicht gefundenen Dokumente dann doch zu bekommen. Man kann erst einfach die

Anfrage formulieren und lässt das System die optimale Anfrage weiter finden. Außerdem wird die gesamte Zeit verkürzt.

2.4.1.1 Relevanzfeedback

Eine bekannte Methode zur neuen Anfrageformulierung ist das Relevanzfeedback, bei dem der Nutzer die *r* Dokumente aus der Ausgangliste danach bewertet, welche Dokumente für ihn relevant sind und welche nicht. Das Verfahren des Relevanzfeedbacks beinhaltet zwei wesentliche Techniken, und zwar die Anfrage expandieren und die Ausdrücke wieder gewichten.

Der Vorteil des Nutzerrelevanzfeedbacks ist größer als bei anderen Anfrageerweiterungen,

1. weil relevante Dokumente vom Nutzer direkt gewählt werden und die ausgewählten Dokumente dieselben suchenden Konzepte haben, die die Anfrage repräsentieren können und nicht in der Anfrage gefunden werden.
2. weil das Durchsuchen in kleine Stücke zerlegt wird, um es einfacher zu gestalten.
3. weil einige Ausdrücke durch das Augenmaß vom Nutzer betont werden und einige nicht.

Die Anfrageerweiterung und die Ausdrucksneugewichtung für das Vektormodell

Das Ähnlichkeitsmaß zwischen der Anfrage und den Dokumenten in der Sammlung beispielweise durch das Vektormodell, $sim(d_j, q)$, kann man etwas verbessern, indem die Anfrage q verfeinert wird, um ein höheres Ähnlichkeitsmaß zu erreichen. Bei diesem Verfahren kennt man in der Wirklichkeit die Menge C_r aber zunächst nicht. Um die Methode des Relevanzfeedbacks anzupassen, wird die Menge der Dokumente D_r, die durch den Nutzer aus der Ausgangliste als relevant gekennzeichnet wurden, statt der Menge der realen relevanten Dokumente C_r definiert. Die Menge der irrelevanten Dokumente D_n wird auch von der Nutzerkennzeichnung bestimmt. Wenn die Mengenfunktion, $|\bullet|$, die Anzahl der Mitglieder wiedergibt, kann man die Relation der Anfragemodifizierung durch einen spezialen Parameter berechen (siehe Formel 20; Anhang 7.5). Die bekannte Anfragemodifizierung ist von Rochio (siehe Formel 21; Anhang 7.5).

Die Vorteile sind ganz klar, da es einfach ist und ein gutes Ergebnis erreicht werden kann. Aber der Nachteil ist, dass es keine optimalen Kriterien gibt.

Evaluierung des Relevanzfeedbacks

Weil die Methode des Nutzerrelevanzfeedbacks durch den Nutzer unterbrochen wird und die relevanten und irrelevanten Dokumente durch den Nutzer gekennzeichnet werden, kann das konventionelle Recall-Precision-Vergleichsmaß nicht geeignet genutzt werden.

Um die Leistung des Nutzerrelevanzfeedbacks einzuschätzen, kann man diese nur abhängig von der restlichen Sammlung berechnen. Die restliche Sammlung ist die Menge derjenigen Dokumente, die nicht in die Menge des Relevanzfeedbacks fällt. Anhand dieser Methode kann man die Leistungsfähigkeit zwischen unterschiedlichen Nutzerrelevanzfeedbackstrategien vergleichen. Die Bewertung von Recall und Precision ist relativ zu der restlichen Sammlung.

2.4.1.2 Lokale Kontextanalyse

Die Erweiterung der Anfrage hilft dem Nutzer, seine Anfrage dem geeigneten Index in der Sammlung anzunähern. Die Wörter in der originalen Anfrage tragen wahrscheinlich die gleiche semantische Bedeutung wie die in der Indexierung, aber eventuell werden sie durch unterschiedliche Ausdrücke beschrieben. Anhand der originalen Anfrage ist dies vielleicht nicht genug, um die benötigten Dokumente zu bekommen. Die Anfrageerweiterung durch eine Lokalanalyse oder eine Globalanalyse ist eine Möglichkeit, um das bessere Ergebnis zu erreichen.

Im Vergleich zur Globalanalyse wird man die erweiterte Anfrage unter der lokalen Strategie zur Laufzeit des Anfrageprozesses erhalten, während die Thesauri unter der Globalanalyse vor dem Suchprozess erzeugt werden müssen. Bei der lokalen Kontextanalyse werden die lokale Analyse und Globalanalyse zusammengesetzt. Die Mehrwortgruppen, die vielleicht einzelne Nomen sind oder aus mehreren Wortarten bestehen, werden als Konzepte (siehe Abschnitt 2.5.1.3) des Dokumentes repräsentiert. Diese Konzepte werden durch den Wert zwi-

schen ihrem Zusammentreffen und der Anfrage ausgewählt. Dieser Wert für die Häufigkeit des Zusammentreffens entstammt eigentlich der Methode der Globalanalyse.

Die drei Schritte der lokalen Analyse können folgendermaßen definiert werden.

1. Während die Anfrage eingegeben wird, werden die Dokumente durch den Retrievalalgorithmus eingeordnet. Inzwischen werden die Dokumente durch das festgesetzte Fenster aufgespaltet. Jeder abgetrennte Teil wird „Passage“ genannt.
2. Jedes Konzept *K* von den ersten *r* Passagen in der Liste wird genommen. Die Ähnlichkeit zwischen der Anfrage und dem Konzept wird durch tf-idf berechnet (siehe Formel 25; Anhang 7.5).
3. Die ersten m eingeordneten Konzepte von 2. Schritt werden zu der Anfrage addiert, um ihr ein neues Gewicht zu geben. Das neue Gewicht wird mit dem Wert $1-0{,}9\times\frac{i}{m}$ addiert, wobei *i* der Platz des Konzeptes von den *m* Plätzen der Rangliste ist.

Durch das oben genannte Ähnlichkeitsmaß $Sim(q,K)$ wurde schon bewiesen, dass dieses Formular gut zu der TREC[3] Dokumentsammlung passt. Aber dieses Formular passt nicht gut zu anderen Dokumentsammlungen. Dadurch sollen die Einstellungen für das Formular abhängig von der Dokumentensammlung angepasst werden.

2.4.1.3 Ähnlichkeitsthesaurus

Ein Ähnlichkeitsthesaurus entsteht durch die Globalanalyse der Dokumentsammlung. Bei der globalen Strategie werden alle Dokumente in der Sammlung betrachtet, damit eine globale ähnliche Thesaurusstruktur erschaffen werden kann. Die erzeugten Thesauri werden später zur Anfrageerweiterung genutzt.

[3] Text REtrieval Conference (siehe http://trec.nist.gov/)

Der Ähnlichkeitsthesaurus wird von der Relation zwischen Ausdrücke aufgebaut. Die Ausdrücke werden in den Semantikraum eingefügt. Unter diesem Blickwinkel werden die Ausdrücke in den Dokumenten, in denen sie auftauchen, indexiert.

Man kann die Relation zwischen den Indexierungsausdrücken k_u und k_v durch den Korrelationsfaktor aufbauen (siehe Formel 26; Anhang 7.5).

Das Bemerkenswerte ist, dass der oben genannte Korrelationsfaktor für alle Dokumente in der Sammlung berechnet wird. Dieser Korrelationsfaktor kann die globalen Ähnlichkeitsthesauri der Sammlung repräsentieren. Weil die Zeitkosten der Berechnung sehr hoch sind, wird dies aber nur ein Mal berechnet. Der Korrelationsfaktor kann auch später aktualisiert werden, wenn neue Dokumente in die Sammlung eingefügt werden.

Wenn man die globalen Ähnlichkeitsthesauri hat, kann man die Anfrage durch diese globalen Ähnlichkeitsthesauri wie folgt erweitern.

1. Die Anfrage muss in den Semantikraum eingefügt werden. Die Funktion wird durch die Formel 29 (siehe Anhang 7.5)definiert.
2. Die Ähnlichkeit zwischen der Anfrage und den Indexausdrücken wird durch die Formel 30 (siehe Anhang 7.5) berechnet.

   In diesem Schritt wird die Korrelation zwischen allen Semantikraum k_u in den Anfrageausdrücken q mit den anderen Semantikraum k_v durch den Ähnlichkeitsthesaurus berechnet.
3. Beliebige Konzeptausdrücke k_v werden mit Hilfe der Ähnlichkeit $Sim(q, k_v)$ eingeordnet, die im zweiten Schritt berechnet wurde. Die ersten r Ausdrücke in der Rangliste werden zu der originalen Anfrage q addiert. Wir werden eine neue erweiterte Anfrage q' bekommen. Das Gewicht $w_{v,q}$ wird berechnet (siehe Formel 31; Anhang 7.5).

Die erweiterte Anfrage q' wird für die neue Anfrage des Retrievalsystems verwendet. Anhand der Erweiterung erhält man ein im Vergleich zu der originalen Anfrage unterschiedliches Suchergebnis.

Die andere einfache Methode ist die Nachahmung des Vektorraummodells. Die Ausgangsdokumente werden unter dem folgenden Ähnlichkeitsmaß eingeordnet (siehe Formel 32; Anhang 7.5).

Diese Methode basiert auf der Idee der Relation zwischen den Ausdrücken und den Konzepten. Ein Teil von dieser IR-Hilfstechnik gehört zum Dokumentenverfahren, indem alle Dokumente in der Sammlung komplett bearbeitet werden, um die Relationsmatrix zwischen den Ausdrücken zu erzeugen. Aber die Relationsmatrix wird zur Laufzeit aber für die Anfrageerweiterung verwendet. Dadurch kann man auch sagen, dass der Ähnlichkeitsthesaurus zum Anfrageverfahren gehört.

2.4.1.4 Thesauri

Die Thesauri könnten auf den ersten Blick als Unterstützung des Problems betrachtet werden, wenn zwei unterschiedliche Wörter bzw. Bezeichnungen für das gleiche Konzept verwendet werden, z.B. sagt einer „Besprechung“ aber der andere sagt „Treffen“. Die Anfrage vom Nutzer kann vielleicht nicht genau auf einen Indexausdruck stoßen. Es ist nicht gemeint, dass die Dokumentsammlung keine benötigte Information beinhaltet. Durch die Thesauri wird die Anfrage sowohl mit den anderen Bezeichnungen als auch in der Originalform geleistet. Dies würde die Wahrscheinlichkeit erhöhen, die gesuchten Dokumente zu bekommen.

Der Thesaurus stützt sich auf die vorübersetzte Liste der wichtigen Wörter, die sich in unterschiedliche Fachgebiete aufteilen, und die Wörter in der Liste, die als verwandte Wörter eine Rolle spielen. Den Thesaurus im IR-System zu verwenden, hat folgende Vorteile:

- Der standardmäßige Wortschatz für den Index und das Suchen wird genutzt.
- Die Anfrage vom Nutzer wird durch einen Ähnlichkeitsthesaurus verbessert.
- Die klassifizierte Hierarchie für die locker und eng verwandten Bedeutungen der aktuellen Anfrage wird genutzt.

Der Nutzen des Thesaurus basiert oft auf kontrollierten Schlagwörtern, da die Schlagwörter die semantische Bedeutung klar hervorheben und das Retrieval eher auf dem Konzept als auf den einzelnen Wörtern basieren. Aus Sicht des IR besteht der Thesaurus aus dem Schlagwort, den verwandten Wörtern und dem Entwurf der verwandten Ausdrücke.

Um die Thesauri aufzubauen, kann man entweder einer automatische oder manuelle Methode verwenden. Durch die meisten manuellen Thesauri sind bloße Generalwörter abgedeckt. Da der Aufbau der manuellen Thesauri viel Arbeitzeit braucht, entstehen kaum manuelle Thesauri für Fachgebiete. Es folgt eine kurze Beschreibung von automatischen Techniken, um Thesauri aufzubauen (vgl. [GRFR98]).

Das Zusammentreffen der Ausdrücke

Alle Ausdrücke werden in einen Vektor umgeformt und durch den Ähnlichkeitskoeffizienten im Euklidischen Raum verglichen. Die lockere Relation zwischen den Wörtern wird vom Vorkommen des Wortpaares erzeugt. Das Resultat ist die Ähnlichkeitsmatrix zwischen den Ausdrücken. Ein asymmetrischer Ähnlichkeitskoeffizient differenziert die Relation zwischen dem Ausdruck *i* und Ausdruck *j*, d.h. $SC(t_i, t_j) \neq SC(t_j, t_i)$ wobei $SC(A, B)$ ein Ähnlichkeitskoeffizient ist.

Kontextausdruck

Bei diesem Ansatz wird der Kontext in der Umgebung jedes Ausdruckes betrachtet, um den den Ausdruck repräsentierenden Vektor zu erzeugen. Durch die Betrachtung der Kontextumgebung kann eine Bedeutung des Ausdruckes aufgefunden werden, weil die semantische Bedeutung des mehrdeutigen Wortes bzw. des Mehrzweckwortes von ihrer Umgebung abhängig ist.

Cluster mit Singularwertzerlegung

Um die Thesauri mit dieser Strategie aufzubauen, stehen zwei Techniken zur Verfügung. Erstens werden drei Matrizen erzeugt, nämlich die Matrix der term-term-Kookkurrenz, die Matrix des Zusammentreffens der *p* höchsten anzutreffenden Ausdrücke zwischen gefundenen Clustern und die Matrix für alle Ausdrücke in der Sammlung. Die letzte Matrix wird zerlegt

und der Singularwert wird berechnet. Die zweite Technik ist, dass die Vektormatrix des Kontextes verwendet wird, um die Anfrage bezüglich ihres Kontextvektors in Cluster zu verpacken. Die Anfrage wird in drei getrennte Cluster geteilt. Die Anfrage wird in jedem Cluster getrennt durchgeführt. Die Ergebnisse nach der Durchführung von jedem Prozess werden zusammenrechnet, um die Dokumente einzuordnen. Die erhaltenen Dokumente sind relevant mit der Anfrage und beziehen sich auf allen Einträgen der Anfrage.

Nur Dokument in Cluster verpacken um ein Thesaurus zu erzeugen

Zunächst werden die Dokumente in der Sammlung durch den Cluster-Algorithmus verteilt. Die Ausdrücke von jedem Cluster werden durch einen geeigneten Auswahloperator herausgenommen, um der Vertreter des Clusters zu sein, die sogenannte „Thesaurusklasse“. Die Anfrage wird durch die aufgebaute Thesaurusklasse expandiert.

2.4.1.5 Semantiknetz

Das Semantiknetz ist das Netz der Begriffe, in dem der Knoten einen Begriff repräsentiert und der Bogen bzw. die Verbindung, die Beziehung zwischen den Begriffen vertritt. Die Information des Knotens identifiziert den individuellen Charakter des Begriffes, den sogenannten Frame. Jeder individuelle Eintrag wird Steckplatz genannt. Obwohl das Semantiknetz im IR-System zwecks Verbesserung verwendet wird, war das Ergebnis noch nicht gut; auch ist das Semantiknetz von der Sprache abhängig.

Das Semantiknetz hat das gleiche Ziel wie die anderen Anfrageerweiterungen, und zwar, das Problem der Fehlanpassung zu lösen. Die Ersetzung des Ausdrucks bzw. Konzeptes mit anderen gleichen semantischen Wörtern erhöht die Chance, um die benötigten Dokumente zu erreichen, oder vielleicht den Weg abzulenken. Die semantische Distanz zwischen zwei Ausdrücke im Netz wird gemessen, um die Beziehung zwischen Ausdrücken abzuschätzen. Dies wird zu den relevanten Dokumenten verwiesen.

Der Thesaurus und das Semantiknetz sind quasi ähnlich. Der Unterschied besteht darin, dass das Semantiknetz weitere komplizierte Informationen repräsentieren kann, z.B. IS-A Hierarchie.

2.4.2 Dokumentverfahren

Bei der schriftlichen Sprache wird der Größeteil der Bedeutung normalerweise durch die Nomen getragen. Deshalb dient das Nomen als Indexierungsausdruck. Obwohl die Bedeutung durch das Nomen getragen wird, gibt es auch andere bedeutende Wortarten, die die semantische Bedeutung vollenden und mit dem Nomen in der Anfrage verwendet werden.

Beim konventionellen Informationsretrieval wird das Dokumentverfahren verwendet, um die unnötigen Wörter zu beseitigen. Gleichzeitig bleiben die wichtigen, bedeutenden Wörter nach dem Verfahren noch weiter erhalten. Die übrig gebliebenen Wörter werden ausgewählt, um die Indexierung zu sein. Mit den übernommenen Wörtern entsteht das Rauschen in der Retrievalanwendung. Im Gegenteil dazu wird deutlich, dass durch eine ungenügende Indexierung die Leistung des Retrievals verringert wird. Das heißt, dass entweder die übernommenen unnötigen Dokumente auftauchen oder die benötigten Dokumente nicht vorkommen.

2.4.2.1 Lexikonanalyse

Die Lexikonanalyse ist der Umwandlungsprozess, in dem die Kette der Zeichen zur Kette der Wörter umgewandelt wird. In den meisten Sprachen trennt man die Wörter einfach durch eine Leerstelle. Aber es gibt auch Sprachen, bei denen man ohne Leerstelle schreibt, beispielweise thailändisch, chinesisch. Um die Wörter in der fortgeschriebenen Sprache zu trennen, muss der Trennungsalgorithmus in dem Lexikonanalyseverfahren verwendet werden.

Außer der Worterkennung gibt es noch andere wichtige Merkmale: die Zahlen, der Trennstrich und das Interpunktionszeichen bzw. Satzzeichen. Die meisten Zahlen können nicht zur Indexierung verwendet werden, weil sie keine klare Bedeutung tragen. Aber es gibt einige Zahlen, die als Indexierung verwendet werden können, beispielweise die Bankleitzahl von einer Bank oder die Postleitzahl.

Die meisten Trennstriche werden verwendet, um Wörter zu verbinden. Ein Wort mit einem Trennstrich hat vielleicht eine neue Bedeutung, beispielweise third-party oder knick-knack. Aber es gibt noch einige Trennstriche, die aufgrund unterschiedlicher Schreibweisen verwendet werden, wo aber die Bedeutung nicht unterschiedlich ist.

Das auffällige Interpunktionszeichen, das in einem Text vorkommt, wird bei der Abkürzung verwendet, beispielsweise zzgl., a.m. (ante meridiem), a.M. (am Main).

2.4.2.2 Stoppwörter weglassen

Die Wörter, die eine hohe Häufigkeit haben, sind für den Retrievalzweck nutzlos. Solche Wörter werden als Stoppwörter bezeichnet. Stoppwörter sind nicht nur Artikel, sondern auch Präpositionen, Konjunktionen und andere Wortarten in der Dokumentsammlung.

2.4.2.3 Stammformen

Ein Wort kann meist in vielen Formen auftreten, beispielsweise als die Pluralform eines Nomens oder als die Vergangenheitsform eines Verbs. Insbesondere in der deutschen Sprache gibt es viele abgeleitete Wörter wie z.B. die Deklination und die Konjugation.

Die Strategie der Stammformreduzierung wurde entwickelt, um die Variantenformen zu reduzieren. Diese Strategie basiert auf der Affixausräumung, dem Nachsehen in Tabellen, der Nachfolgeabwechslung und der n-gramm-Methode. Um das Suffix zu reduzieren, wird meistens für den englischen Retrievalzweck der Algorithmus von Porter verwendet.

2.4.2.4 Indexausdrücke wählen

Zwei Möglichkeiten, Wörter als Index zu wählen, sind zum einen die Auswahl aller Wörter (Volltext) und zum anderen die Auswahl einiger Wörter. In der bibliografischen Wissenschaft wird die Auswahl durch einen Experten bearbeitet. Die automatische Auswahl ist eine bekannte Methode für das Informationsretrieval.

Weil das Nomen und die Nomengruppe bzw. Mehrwortgruppe immer die beste Semantik liefern, werden die Nomen bei der automatischen Indexierung hauptsächlich ausgewählt. In der englischen Sprache können mehrere Nomen eine singularische Komponente definieren, beispielweise „night train“. Aber das vorige Beispiel kann nur ein Wort in der deutschen Sprache sein, und zwar „Nachtzug“. Das ist ein Unterschied zwischen den Sprachen.

Außer dem obengenannten Beispiel gibt es noch einen anderen Fall, in dem das Adjektiv die Semantik des Nomens vollendet, beispielweise auf Englisch „west europe“. Wenn wir im Wörterbuch nachschlagen, finden wir, dass das englische Wort „west“ das Nomen sowie auch das Adjektiv oder Adverb sein kann. Im Vergleich zu der englischen Sprache heißt das deutsche Wort „Westeuropa“. Dieses Problem behindert das krosslinguale Informationsretrieval (siehe Kapitel 4).

Zumindest sind die Nomengruppen in der englischen Sprache für eine bestimmte semantische Bedeutung wichtig. Aus diesem Grund werden die Nomengruppen verwendet, um Indexe zu repräsentieren. Die einfachste und praktische Methode zur Erkennung der Mehrwortgruppe ist der definierte Abstand zwischen den Nomen.

2.4.2.5 Syntaxanalyse

Beim Informationsretrieval wird der Text normalerweise als ein Wortbehälter betrachtet. Dadurch wird die semantische Bedeutung verloren. Der Einzelausdruck-Ansatz nutzt beim Retrievalprozess die Phrasen im Dokument. Die Phrase besitzt im Wortbehälter eine Bedeutung. Der hoch entwickelte Ansatz zur Ermittlung der Phrase basiert auf dem Algorithmus, der allgemein für die natürlichere Sprachverarbeitung verwendet wird. Die Ansätze schließen Wortklassemarkierungen[4], syntaxgrammatische Definitionen und Informationsextraktionsheuristik ein.

Einzelner Ausdruck

Der Einzelausdruckansatz ist lange bekannt, weil er sehr einfach zu implementieren ist. Der Einzelnausdruckansatz hilft dem Nutzer, um seine Anfrage den Ausdrücken im Dokument anzupassen. Das beträchtliche Problem bei diesem Ansatz ist der Unterschied in der Großschreibung, des Präfixes und des Suffixes. Außerdem muss man auch das Problem des überflüssigen Stoppwortes und das der besonderen Zeichen betrachten, z.B. der Bindestrich, das Hochkomma, das Komma usw. Bei englischer Sprache gibt es einen Algorithmus oder eine

[4] Part-of-Speech Tagger

Initialregel, um die meisten Probleme zu lösen, z.B. den Algorithmus von Porter und von Lovins zur Stammformreduzierung – (Präfix, Suffix). Einige Methoden des Einzelausdruckes werden in der Lexikonanalyse (siehe Abschnitt 2.4.2.1) weiter beschrieben.

Einfache Phrase

Die einfache Phrase wird von einem Paar der Ausdrücke identifiziert. Zwischen einem Paar darf kein Stoppwort gefunden werden und es muss (laut dem Bericht von TREC in [GRFR98]) mehr als 25 mal getroffen werden. Als Ergebnis wird berichtet, dass die Phrase die Precision und Recall positiv beeinflussen kann. Insofern könnte man die einfache Phrase in Betracht ziehen, sie vertritt die Ausdrücke aber nicht.

Komplizierte Phrase

Ein Problem, das die IR-Entwicklung verhindert, ist die in einem Dokument gefundene Uneindeutigkeit. Insofern wird die natürliche Sprachverarbeitung in dem IR-System aufgenommen. Die natürliche Sprachverarbeitung beschäftigt sich mit dem Aufbau der kanonischen Struktur, um die Bedeutung zu begreifen.

Obwohl viele Hilfeanwendungen der natürlichen Sprachverarbeitung noch nicht völlig funktionieren, um die Sprache zu verstehen, helfen sie aber das IR-System genügend zu verbessern, insbesondere die Phrase im Dokument zu identifizieren.

Wortklasse-Tagging und Wortsinne-Tagging

Die Tagging-Arten teilen sich in zwei Gruppen, nämlich Wortklasse-Tagging bzw. POS-Tagging und Wortsinne-Tagging. Wortklasse-Tagger basiert auf entweder der Statistik- oder Regelbasismethode. Das einzige Ziel ist, den Text in kleine Sinneinheiten zu teilen und die Wortklasse zu identifizieren. Wortsinne-Tagger verwendet die auf Wörterbuch basierende Stammform, um den Sinn des Wortes zu identifizieren.

Syntaktische Phrase

Bei diesem Ansatz geht es darum, dass die Wörter in einem Satz als ihre syntaktische Art identifiziert werden, z.B. Nominativ, Verb, Akkusativ, usw. [GRFR98] berichtet von TREC,

dass die Dateimenge von TREC-5, auf der die Indexe auf der Stammform, einfacher Phrase, dem Köpfergänzungspaar und dem Personnamen basiert, im Durchschnitt eine Verbesserung von 20 % der Precision erreichte. Während die Laufzeit des Systems dramatisch erhöht wird, lohnt sich die Leistung durch die syntaktische Phrase nicht.

2.4.2.6 Informationsextrahierung

Informationen über den Personennamen, dem Ort, der Institution, usw., könnten den Nutzer bei seiner Anfrage unterstützen. Die Suchmethode in einer der solchen strukturierten Datei in unstrukturierten Dokumenten ist auf Informationsextrahierung fokussiert. Das IR-System könnte diese Extrahierung durch die Gewichtzunahme für extrahierte Ausdrücke einschließen.

2.4.3 Anfrageprozess beschleunigen

Während die Retrievalmethoden und die IR-Hilfstechnik sich mit der Suche der relevanten Dokumente bezüglich einer Anfrage beschäftigen, wird in diesem Abschnitt beschrieben, wie die Begriffe von den Dokumenten repräsentiert werden und wie schnell die relevante Position in der Dokumentsammlung ermittelt wird. Außerdem wird die Größe des Speichers betrachtet.

Anstatt dass die Anfrage mit den Dokumenten direkt berechnet wird, kann sie mit den vertretenden Wörtern der Dokumente ermittelt werden. Die vertretenden Wörter werden „die Indexe“ genannt. Das als Index dienende Wort wird mit der echten Lage im Dokument verbunden. Zur Indexierung wurden viele Methoden entwickelt. Die wichtigen Betrachtungen sind der Arbeitspeicher und die Suchzeit, weil auf die benötigten Dokumente im IR-Prozess immer über den Index zugegriffen wird. Natürlich braucht die große Indexmenge viel Platz. Mit einer guten Indexstruktur kann der Arbeitspeicher geschont und die Zugriffzeit verkürzt werden.

Einige Indexmethoden brauchen viel Arbeitspeicher, aber wenig Zeit zum Zugriff, z.B. der Suffixbaum. Einige brauchen nicht viel Platz, aber längere Zeit zum Zugreifen, z.B. das sequentielle Suchen oder die Knuth-Morris-Pratt-Methode. Die bekannteste Indexmethode ist

die invertierte Liste, weil bei dieser Methode die Platzkomplexität nur $O(n^{\gamma})$ bis $O(n)$ und die Zeitkomplexität nur $O(m \log n)$ bis ca. $O(n)$ betragen, wobei n die Größe der Datenbank und m die Länge des Suchwortes ist. Die andere gute Methode ist der Boyer-Moore-Algorithmus, der als Platzkomplexität nur $O(m+\sigma)$ und als durchschnittliche Zeitkomplexität nur $O(n \log(m)/m)$ hat.

2.4.3.1 Invertierte Liste

Die invertierte Liste ist ein Wortorientierungsmechanismus für die Schlagwörter in der Dokumentsammlung, um den Suchprozess zu beschleunigen. Die invertierte Liste besteht aus dem Wörterverzeichnis und ihrem Vorkommen. Das Wörterverzeichnis ist die Liste der Schlagwörter. Das Vorkommen ist die Stelle, an der das Schlagwort vorkommt. Diese Stelle wird durch Wort-, Charakter- oder Blockadresse angezeigt.

Die invertierte Liste besteht aus zwei Komponenten, nämlich die Liste aller Ausdrücke, der sogenannte „Index“, und die Menge der Positionsliste, sogenannte „Positionsliste“. In einer Positionsliste stehen alle unterschiedliche Ausdrücke in Form des Tupels, $\langle doc\_id, tf \rangle$.

Für eine Wort- oder Charakteradresse wird viel zu viel Speicherplatz für den Prozess verbraucht, weil jedes Wort oder jeder Charakter eine eigene Adresse in der Struktur verlangt. Die Blockadresse braucht nicht so viel Platz, um auf die Stelle zu zeigen, aber sie braucht mehr Zeit, um die richtigen Ausdrücke in der Blockadresse zu suchen.

Ein typischer dekomprimierter Index benutzt 4 Byte für die Bezeichnung des Dokumentes (*doc_id*) und zwei Byte für die Häufigkeit des Ausdruckes (*tf*) (vgl. [GRFR98]). Drei Dateien werden bei invertierter Liste produziert,

- Indexdatei : in der Indexdatei ist die Positionsliste beinhaltet. Jeder Ausdruck in der Sammlung wird in Form $t_j \rightarrow (d_1, tf_{1j}), (d_2, tf_{2j}), ..., (d_i tf_{ij})$ geschrieben, wobei d_i das Dokument *i* und tf_{ij} die Häufigkeit des Ausdruckes repräsentiert.
- Dokumentsdatei : in der Dokumentsdatei ist die Information jedes Dokumentes beinhaltet.

- Gewichtsdatei : in der Gewichtsdatei ist das Gewicht aller Dokumente in Form $d_i = (w_{i1}, w_{i2}, ..., w_{ij}, ..., w_{in})$ beinhaltet, wobei w_{ij} das Gewicht zwischen dem Dokument *i* und dem Ausdruck *j* von *n* Ausdrücken ist.

Den Suchprozess von der invertierten Liste kann man in drei Schritte teilen:

1. Die Schlagwörter suchen.
2. Ihre Stelle erhalten.
3. Die Wörteradresse manipulieren.

Für eine große Textsammlung ist es vernünftig, die Wörterverzeichnisdatei von der Adresse abzutrennen. Diese abgetrennte Datei kann auf dem Hauptarbeitsspeicher des Rechners resident erhalten werden. Der komprimierte Index kann den benötigen Speicher bis zu 10% vom originalen Text verringern. Die Größe der Positionsliste für den invertierten Index kann durch die Zipf-Regel abgeschätzt werden. Die passende Dateistruktur hilft, den Wörterzugriff schneller zu erreichen. Die bekannten Dateistrukturen sind die Hashfunktion sowie die Baum- und die Binärbaumstruktur.

Bei der Hashfunktion und der Baumstruktur hat der Suchprozess die Zeitkomplexität $O(m)$, wobei *m* die Länge des Suchwortes ist. Die Binärbaumstruktur hat die Komplexität $O(\log n)$ beim Sucheinsatz, wobei *n* die Größe der Datenbank ist.

Nachdem man die Adresse der Schlagwörter erhalten hat, kann man ein Einzelwort genau erfassen. Für die Kontextanfrage werden die Wörter erst einmal geteilt und separat gesucht. Danach lassen sich die Ausgaben mit der separaten Suche verbinden. Die eingeordnete Adresse und die Lagebeziehungstechnik werden schließlich an dieser Stelle verwendet, damit die Übereinstimmung der möglichen Kombination der Wörter eingeschätzt werden kann.

2.4.3.2 Authentifizierende Datei

Eine authentifizierende Datei ist eine Verschlüsselung des Dokumentes. Es muss in zwei Phasen des erforderlichen Retrievals implementiert werden. Es geht bei der ersten Phase darum, dass alle Ausdrücke der authentifizierenden Datei durchgesucht werden und die getroffenen

Ausdrücke identifiziert werden. In derzweiten Phase wird der Originaltext des Dokumentes mit der erhaltenen Ergebnisliste noch mal durchsucht, um sicherzustellen, dass die getroffenen Ausdrücke richtig sind.

Jedes Wort wird von einer Hashfunktion verschlüsselt. Bei der Verschlüsselung wird meist die Menge eines Bit des Authentifikators benutzt. Bei der Abgleichung wird vielleicht ein Fehler auftreten, wenn der Ausdruck nicht in der Liste des Authentifikators liegt aber er die gleiche Bitmap wie eine in dem Authentifikator hat. Im Bericht von Standfill und Thau wird die Chance der falschen Abgleichung für den Authentifikator mit der Größe 1.024 Bit unter drei Prozent [GRFR98] angegeben. Im Abgleichungsprozess wird die authentifizierende Anfrage mit allen Dokumenten durch die boolesche Logik „und“ durchgeführt. Der boolesche Authentifikator kann aber die Information der Umgebung und die Information des Gewichts von dem in einem Dokument aufgetauchtem Ausdruck nicht abspeichern.

2.4.4 Mensch-Maschine-Schnittstelle

Während der Laufzeit des automatischen Systems kann man nicht genau vorhersagen, was man zum Schluss bekommt oder ob alles von dem Suchergebnis auch die gesuchten Informationen ist. Es ist eine Tatsache, dass es einen Kenntnisabstand zwischen dem Nutzer und dem Systemsprozess gibt. Wenn man diesen Prozess verfolgen könnte, würde dieser Abstand verringert und der Grund des Abstands wahrscheinlich aufgefunden. Die Suchmethode mit Anfragenavigator könnte einem Nutzer helfen, um seine gesuchte Information herauszufinden [GRWE02].

Weil die meisten IR-Nutzer ihren persönlichen Verstand zum Retrieval der benötigten Information einsetzen, hilft eine gute Mensch-Maschine-Schnittstelle solchen Nutzern, um den Verstand durch das System zu unterstützen sowie die gesuchte Information zu beschreiben. Die Hilfe der Mensch-Maschine-Schnittstelle besteht auf der Formulierung der Anfrage, der Auswahl der Informationsquelle, dem Verständnis des Suchergebnisses und dem Verfolgen des Suchablaufs.

In der Praxis kann der Nutzer nicht genau wissen, welchem Fachgebiet in den Korpora seine Anfrage entspricht. Mancher Nutzer kennt nicht den Wortschatz im Fachgebiet, aus dem er gerade Informationen benötigt. Er weiß wahrscheinlich nicht, wie das Gewicht zwischen den Ausdrücken und der Anfrage oder den Ausdrücken und den Dokumenten funktioniert[GRWE02]. Dabei hilft die Mensch-Maschine-Schnittstelle dem Nutzer, damit er den zu suchenden thematischen Sinn erfassen und die benötigte Information durch die empfohlenen Ausdrücke des Systems erschließen kann.

2.4.4.1 Prinzipieller Entwurf

Um eine gute Mensch-Maschine-Schnittstelle zu entwerfen, muss man zuvorderst folgende Eigenschaften bedenken.

Angebot des informativen Feedbacks

Ein guter Entwurf der Mensch-Maschine-Schnittstelle sollte sich mit den Fragen beschäftigen, wie der Entwurf den Nutzer mit dem zugehörigen Feedback versorgt,

- die Beziehung zwischen der Anfragevorschrift und den erhaltenen Dokumente
- die Beziehung zwischen den erhaltenen Dokumenten
- die Beziehung zwischen den erhaltenen Dokumenten und beschreibenden Daten zu den Sammlungen.

Wenn der Nutzer die obengenannten Fragen beantworten kann, hat er die lokale Kontrolle über das System.

Verringerung der laufenden Speicherbelastung

Bei manchem IR-System muss man viel Geduld haben, um die unbenötigten Dokumente aus der Rangliste herauszufiltern. Diese Tatsache hat viele Ursachen. Unter anderem die nicht angemessene Anfrage. Wenn der Nutzer die unnötigen Wörter bzw. Attribute durch die Mensch-Maschine-Schnittstelle entfernen und nur die benötigten Konzepte zurückbehalten kann, werden der Suchprozess und die laufende Speicherbelastung reduziert.

Versorgung des alternativen Interfaces für die Anfänger und Experte

Weil es viele verschiedene Nutzer gibt, sollte die flexible Mensch-Maschine-Schnittstelle für allgemeine Nutzer angepasst werden, nicht nur für den anfänglichen Nutzer, sondern auch für den Experten. Unter einem guten Entwurf der Mensch-Maschine-Schnittstelle sollte die einfache Form für den Anfänger verwendet werden, und der Experte sollte die Interaktion passend zur angewandten Suche einstellen dürfen.

2.4.4.2 Evaluierung des Interaktionssystems

Die Mensch-Maschine-Schnittstelle im Informationsretrieval verursacht eine Abhängigkeit der erhaltenen Dokumente, da die Ausgangsdokumente von der Entscheidung des Nutzers abhängig sind. Der Unterschied in der sprachlichen Fähigkeiten, der Vernunft und der Persönlichkeit von verschiedenen Nutzern beeinflusst den Unterschied des Suchergebnisses durch die Mensch-Maschine-Schnittstelle. Obwohl die innovative Mensch-Maschine-Schnittstelle einem anfänglichen Nutzer gut hilft, hindert die Entwicklung vielleicht den Experten.

Weil ein Teil des Suchprozesses von der Entscheidung des Nutzers abhängig ist, sind die Precision- und Recallwerte nicht geeignet, um das Können des interaktiven Systems zu messen. Eine Möglichkeit zur Einschätzung der Leistung des interaktiven Systems ist, dass die Precisionwerte, R-Precision (siehe Abschnitt 2.3.2.1), an den ersten r Stellen der Ergebnisliste berechnet werden bzw. an einer höheren Recall-Stufe.

Außer Precision und Recall sollte man noch die Lernzeit des Nutzers an dem System, die Dauer von Anfang an bis zum Ergebnis und die Fehlerrate betrachten, damit man verschiedene interaktive Systeme oder ein interaktives System mit einem automatischen System vergleichen kann. Diese Messung ist nicht einfach. Um das interaktive System zu messen, muss man eine psychologische Methode verwenden. Das Ergebnis von der Messung kann man in einem solcherart eingeschränkten Zusammenhang ungefähr ermitteln.

2.4.5 Natürliche Sprachverarbeitung

Bei der Verarbeitung von Sprache auf dem Computer treten Probleme auf, die bei der natürlichen Sprachbenutzung nicht vorkommen. Deswegen muss das Verständnis der natürlichen Sprache zuerst hergestellt werden. Die mathematische bzw. statistische Strategie wird so angepasst, dass Berechungen mit der Sprache ermöglicht werden. Die Reduzierung der sprachlichen Unklarheiten hilft nicht nur dem Computer bei der Zusammenarbeit mit anderen Medien, bessere Leistung zu erbringen, sondern auch dem Nutzer, den Anwendungsprozess leichter zu verstehen und gut mit dem Computer zusammenzuarbeiten. Genau wie bei der Mensch-Maschine-Schnittstelle ist die gesamte Leistung von den Stärken und Schwächen des Interfaces und den pragmatischen Leistungen der Sprache abhängig.

Das Verständnis des sprachlichen Verfahrens kann die Entwicklung der IR-Forschung von Wortebene zu Syntaxebene, Semantikebene und Pragmatikebene verbessern. Dadurch kann die natürliche Sprachverarbeitung für die monolinguale Suche als auch für die krosslinguale Suche sehr behilflich sein, um die Leistung des Rechnens dem sprachlichen Vermögen der Menschen anzunähern.

2.4.5.1 Linguistische Morphologie

In diesem Abschnitt werden einige wichtige Begriffe zum Thema „Morphologie“ erklärt, damit man zunächst die allgemeine Funktion des Wortes grob verstehen kann. Die Wörter befinden sich in verschiedenen Formen, insbesondere bei den europäischen Sprachen. Durch die Aufteilung könnte man den Stamm und die Funktion des Wortes prinzipiell verstehen. Der Prozess, der das Wort von einer Wortform in seine Grundform und die zusätzlichen Teile aufteilt, wird „*morphologische Satz- und Syntaxanalyse*“ genannt.

Die Veränderung des Wortes von der Normalform zur Grundform wird „*Stammformreduzierung*“ genannt, z.B. von „Städte“ zu „Stadt“.

Das *Morphem* bezeichnet meist das minimale Sinngehaltstück in einer Sprache.

Die *Morphologie* ist das Studium davon, wie die Wörter aus den Morphemen aufgebaut werden.

Die Wortformen werden aus der Grundform und zusätzlichen Teilen aufgebaut. Die zusätzlichen Teile heißen *Affixe*. Die Affixe teilen sich in vier Gruppen auf: das Präfix, das Suffix, die Einfügung und das Circumfix[5].

Das Präfix und das Suffix können meist einfach an ein Stammwort angeknüpft werden. Solche Wörter nennt man „*zusammengesetzte Morpheme*“. Es gibt auch einige Sprachen, in denen Wörter durch ein kompliziertes Affix erheblich verändert werden. Bei der komplizierten Ableitungsmethode wird dies „*unzusammengesetztes Morphem*“ genannt.

Die Wortbauweise von den Morphemen kann in zwei Klassen unterteilt werden.

1. Die *Flexion* ist die Aufbaumethode, die den Wortstamm und die grammatischen Morpheme zusammensetzt. Normalerweise sind die Flexionswörter in der gleichen Klasse wie das Stammwort, z.B. die Flexion des Plurals.
2. Die *Derivation* ist die Aufbaumethode, die den Wortstamm und die grammatischen Morpheme zusammensetzt. Das Ergebnis dieser Aufbaumethode ist, dass die abgeleiteten Wörter in anderen Klassen als Stammwörter auftreten. Die abgeleiteten Wörter haben oft eine andere Bedeutung als ihr Stamm. Bei den meisten Derivationswörtern ist es schwierig vorauszusagen, was sie genau bedeuten.

Die Morphologie spielt keine Rolle in einem natürlichsprachlichen Lexikon, insbesondere bei einer produktiven Morphologie wie der Flexion.

Das *Lemma* ist die Menge der lexikalischen Formen, die den gleichen Stamm, den gleichen hauptsächlichen Part-of-Speech und den gleichen Wortsinn haben.

[5] Eine Art von Affix, die ein anderes Morphem mit einem Präfix und einem Suffix umschließt. Das Circumfix tritt im Deutschen häufig auf, z.B. als Partizip II eines schwachen Verbs „gemacht“ („ge“+mach(-en)+“t“) oder z.B. auch bei „unaufhörlich“ („un“+aufhör(-en)+“lich“).

2.4.5.2 Syntaxtische Wortklassen und das Part-of-Speech Tagging

Die Relevanz des Part-of-Speech (POS) für die Sprachverarbeitung ist die Eingabe der allgemeinen Bedeutung der wörtlichen Information entsprechend ihrer Nachbarwörter. Die Definition des POS basiert auf der morphologischen und der syntaktischen Funktion. Die Wörter, die anhand ihrer Affixe auf ihre morphologische Eigenschaft oder anhand ihrer Nachbarwörter auf ihre verteilende Eigenschaft untersucht werden, werden in die Klassen sortiert, die den gleichen oder einen ähnlichen semantischen Zusammenhang haben. Z.B. repräsentieren Adjektive die Eigenschaft des folgenden Nomens. Aber normalerweise wird der semantische Zusammenhang als das Merkmal für das POS verwendet.

Aufgrund der Eigenschaft des POS kann man ihn für viele Sprachverarbeitungsmodelle verwenden, z.B. für die Spracherkennung, die Sprachsynthese, das Informationsretrieval usw. Insbesondere beim Informationsretrieval wird das POS oft eingebracht, um die wichtigen Worttypen zu erkennen. Außerdem wird das POS häufig für die partielle Satzanalyse verwendet, um Namen oder andere nutzbare Phrasen für die Informationsextraktion aufzufinden.

Das POS lässt sich in zwei Unterkategorien aufteilen: die abgeschlossene Klasse und die geöffnete Klasse. Die abgeschlossene Klasse hat relativ feste Zugehörigkeiten, z.B. die Präposition. Der Unterschied liegt darin, dass in der geöffneten Klasse neue oder fremde Wörter vorkommen werden können. Die Wörter in der abgeschlossenen Klasse sind normalerweise Funktionswörter, die meist kurz sind und häufig auftauchen und die grammatikseitig eine große Rolle spielen, wie z.B. of, it, and, or [JUMA00]. Das Nomen, das Verb, das Adjektiv und das Adverb sind die vier wichtigsten Klassen der geöffneten Klasse, wobei es nicht unbedingt alle vier Klassen in jeder Sprache gibt. Die Wörter in der geöffneten Klasse werden auf Englisch „content words“ genannt.

In der Klasse des Nomens gibt es ein großes Hindernis, da die Nomen und die Eigennamen schwer automatisch ohne die Liste der Eigennamen unterschieden werden können. Ein Hindernis liegt bei der Verbklasse nur in der morphologischen Form. Die verschiedenartige Verhältnisweise der Wörter ist bei der Adverbklasse ein Problem, wobei einige Begriffe auch Nomen sein können, z.B. ist das englische Wort „monday“ manchmal ein Zeitadverb.

Das andere Interesse gilt dem Wortpartikel. Das Wortpartikel ist ein Wort, das einer Präposition oder einem Adverb ähnelt und mit einem Verb verknüpft ist, um eine große Einheit zu formulieren, das sogenannte „Phrasal-Verb[6]“. Im Vergleich zur deutschen Sprache kann man das englische Phrasal-Verb mit dem deutschen trennbaren und untrennbaren Verb mit fester Präposition quasi vergleichen.

2.4.5.3 Semantische Variante

Das Problem der sinnlichen Bedeutung des Wortes ist ein Hauptproblem des Informationsretrievals. Dies beeinflusst direkt die Bewertungen Precision und Recall. Dieses Problem lässt sich in vier Gruppe unterteilen.

Die Homonymie

Die Homonymie nimmt Bezug auf die Lexeme[7], die in der gleichen Form sind aber unterschiedliche Bedeutungen haben.

Beispiel von [JUMA00] für das Wort „bank“ im Englischen:

1. “Instead, a bank can hold the investments in a custodial account in the client’s name.”
2. “But as agriculture burgeons on the east bank, the river will shrink even more.”

Die Polysemie

Die Polysemie nimmt Bezug auf die Idee eines einzelnen Lexems mit mehreren relevanten Bedeutungen.

Beispiel von [JUMA00] für Wort „serve“ im Englischen:

1. “They rarely serve red meat, preferring to prepare seafood, poultry or game birds.”

---

[6] *aus einem Verb und weiteren Elementen bestehender Ausdruck* [http://dict.leo.org]

[7] eine fundamentale Grundeinheit im Lexikon einer Sprache

2. “They served as U.S. ambassador to Norway in 1976 and 1977.”
3. “He might have served his time, come out and let an upstanding life.”

Das Synonym

Das Synonym weist den Zusammenhang zwischen unterschiedlichen Lexemen mit gleicher Bedeutung auf.

Beispiel von [JUMA00]:

1. “How big is the plane?”
2. “Would I be flying on a large or small plane?”

Der einfache manuelle Nachweis, welches Wort das Synonym des beobachteten Wortes in einem Satz ist, besteht darin, dass man solche Wörter durch ein mögliches Synonym ersetzen kann und die allgemeine Bedeutung des Satzes betrachtet, ob ihre Bedeutung verändert wird. Falls die allgemeine Bedeutung nicht verändert wird, ist das neue Wort das Synonym des beobachteten Wortes, sonst nicht.

Die Hyponymie

Die Hyponymie weist die Relation zwischen den in den semantischen Oberklassen verwandten Lexemen auf. Die folgende Definition soll dies verdeutlichen.

Wenn x die Hyponymie von y ist und der Satz mit x für irgendeine Situation richtig ist, muss der neue mit y ersetzte Satz an der richtigen Stelle auch richtig sein.

Das ist ein(e) x. → Das ist ein(e) Y.

Zum Beispiel:

Das ist eine Mango. → Das ist ein Obst.

Die Hyponymie und Hypernymie sind der Grund, dass Methoden wie die Ontologie an der IR-Entwicklung beteiligt werden. Der Begriff der Ontologie bezieht sich auf die Menge der ausgeprägten Objekte, die von einem Definitionsbereich analysiert werden. Die Taxonomie

ist das besondere Arrangement der Elemente einer Ontologie als Baumstruktur der semantischen Oberklasse.

2.4.5.4 Natürliche Sprachverarbeitung als Hilfstechnik

Wenn man natürliche Sprachverarbeitung erlangen möchte, kann man die menschliche Sprache als systematische Funktion betrachten. Der Satz, der eine semantische Bedeutung trägt, wird mit Wörtern in einem syntaktischen Zusammenhang konstruiert. Die Syntax und die Wörter spielen dabei eine große Rolle, um die Bedeutung herüberzubringen. Ohne eine festgelegte Syntax versteht man diese kaum, aber man kann sie teilweise erraten.

Die syntaktische Information des Textes wird beim herkömmlichen Informationsretrieval oft ignoriert. Meistens verwendet das Informationsretrieval die sogenannte „der Sack der Wörter"- Methode. Bei dieser Methode werden die wichtigen repräsentativen Wörter von den Dokumenten in einen Sack gepackt, um in den benötigen Dokumenten gesucht oder damit verbunden zu werden.

Die wichtige Verbindung zwischen den Dokumenten in der Datenbank und dem IR-Nutzer sind üblicherweise die Ausdrücke bzw. Indexe, die mit den Dokumenten verbunden werden, und die Suchwörter, die die von einem Nutzer benötigten Informationen repräsentieren. Daher wirken die von einem Nutzer verwendeten Suchwörter direkt auf die Leistung des Systems. Ein großes Problem, nämlich dass viele unbenötigte Dokumente vom Retrievalprozess geliefert werden, stammt wahrscheinlich aus der Unangemessenheit zwischen den Indexen der Dokumente und den Suchwörtern des Nutzers. Der Grund für dieses Problem liegt sowohl in Fehlern beim Schreiben, die in einigen Teilen der Dokumente oder der Suchwörter vorkommen, als auch in unterschiedlichen semantischen Bedeutungen. Außerdem stellen die metaphorischen und metonymischen Ausdrücke ein weiteres Problem dar. Diese entstehen durch Wortspiele des Autors.

Die gleichen Stammwörter haben verschiedene Flexion- und Derivationsformen. Die Frage ist, ob man die Varianteformen als unterschiedliche Wörter einfach stehen lassen oder in die einzelnen Grundformen zerlegen sollte. Wenn man sich nicht für die Grundform interessiert, werden die Varianten als unterschiedliche Wörter mit separater Worthäufigkeit betrachtet.

Der Hauptvorteil der Nutzung von der Stammform ist, dass es einer bestimmte Anfrage bzw. Suchwörtern erlaubt, die in den Dokumenten beinhalteten beliebigen morphologischen Ausdruckvariante zu finden [JUMA00]. Die Nutzung der Stammform erhöht den Recallwert, indem die in den Dokumenten vorkommenden morphologischen Formen entsprechend der Anfrage gefunden werden, während die unrelevanten Dokumente vielleicht auch häufiger auftauchen, d.h. der Precisionwert wird reduziert.

Wenn ein mehrdeutiges Wort wie die Homonymie und die Polysemie als Anfrage verwendet wird, werden die Dokumente, die einen beliebigen Sinn des Wortes beinhalten, wiedergefunden. Die Dokumente, die das Wort mit demselben Sinn wie das vom Nutzer gesuchte beinhalten, werden als relevante Dokumente bewertet, während die, die das Wort mit anderer Bedeutung beinhalten, zu der unrelevanten Gruppe gehören. Folglich bewirken die Korpora, die durch die Homonymie und Polysemie größtenteils abgedeckt werden, einen niedrigen Precisionwert. Ebenso wird der Recallwert reduziert, wenn die Synonyme oder Hyponyme aus den Korpora häufig auftauchen.

Anhand des syntaktischen POS kann das Problem der enormen Anzahl der Polysemie und der Homonymie gelöst werden, indem man eine von zwei fundamentalen Methoden verwendet, nämlich die integrierte Rule-to-Rule-Methode oder die Stand-Alone-Methode [JUMA00].

Eine Anwendung der natürlichen Sprachverarbeitung ist die Verbesserung der Anfrage. Die richtige Verbesserung der Anfrage kann natürlich die Leistung des Systems erhöhen. Eine Möglichkeit zur Verbesserung der Leistung für die Anfrage vom Nutzer ist die Verwendung des auf der Semantikebene basierten Thesaurus. Die Suchwörter, die nicht zu den Wörtern im Index passen, werden durch andere Wörter mit gleicher oder ähnlicher Bedeutung ersetzt. Um den Thesaurus vom Korpus automatisch aufzubauen (vgl. Abschnitt 2.4.1.3 und 2.4.1.4), basiert er meist auf der Korrelation der Wörter statt auf Synonymen. Die bekannteste Methode hierfür ist das Term-Clustering. In dem Thesaurusaufbauprozess, werden die Wörter aus dem Container in ein Cluster gepackt, um die Synonyme zu finden, die zu der Anfrage des Nutzers passen.

Die Ad-hoc-Anwendung ist nicht die einzige auf den Wörtern basierende Aufgabe des Informationsretrievals. Eine andere Aufgabe dieser wörtlichen Basis ist die Kategorisierung der

Dokumente. Die Teilaufgabe der Dokumentenkategorisierung ist das Filtern, z.B. E-Mails akzeptieren oder wegwerfen, Dokumente sammeln, einen Text segmentieren oder einen Text zusammenfassen. Daher spielt die natürliche Sprachverarbeitung für die Begriffsklärung der Wörter eine große Rolle in der Entwicklung einer solchen Anwendung.

Aufgrund dieser Faktoren ist die natürliche Sprachverarbeitung heutzutage in der IR-Forschung ein sehr nützliches Zusatzwerkzeug. Insbesondere für krosslinguales Informationsretrieval kann man nicht auf die natürliche Sprachverarbeitung verzichten. Für die bilinguale Suche durch die SENTRAX[8] wird die Morphologie aus der natürlichen Sprachverarbeitung verwendet. Die Erklärung befindet sich im Kapitel 4.

2.5 SENTRAX für das Informationsretrieval

Die SENTRAX ist eine IR-Anwendung mit einer grafischen Mensch-Maschine-Schnittstelle. Sie bietet vier nützliche Funktionen an. Anhand ihrer Interaktion bzw. Mensch-Maschine-Schnittstelle kann der Nutzer flexibler als mit einer herkömmlichen IR-Anwendung arbeiten. Zwei von vier Funktionen sind die Grafikdarstellungen, wobei sich eine um Tipp- bzw. Schreibfehler kümmert und eine andere die erweiternden Begriffe besorgt. Andere zwei Funktionen sorgt dafür, dass die Ausgangsdokumente in normaler Weise und ähnliche Dokumente erhalten werden.

Die auffälligen Merkmale von der SENTRAX sind die grafischen Funktionen. Eine erste Grafikfunktion zeigt dem Suchenden die variante Schreibweise als Empfehlung (siehe Abschnitt 2.5.2.1). Die im Hintergrund wirksame Technik fußt auf einer SpaCAM. Die SpaCAM-Technik wird in [HEIT94] beschrieben und in [HAGS96]. Eine zweite Grafikfunktion ist die zweidimensionale grafische Mensch-Maschine-Schnittstelle, nämlich die ContextMap (siehe Abschnitt 2.5.2.2), die durch die assoziierten Wörter gefüllt wird. Die auf dem Bildschirm aufgetauchten Wörter haben nicht nur Beziehungen zu der Suchanfrage, sondern auch unter-

[8] Essence Extractor Engine

einander. Diese Beziehungen werden in einer kleinen Gruppe klassifiziert, in der die eng verwandten ausgefilterten Wortarten erfasst werden.

Die grafische Mensch-Maschine-Schnittstelle der SENTRAX stellt ihren grafischen Dienst für den Nutzer mit den untergliederten Wörtern dar, die mit seiner Anfrage verwandt sind. Anhand der aus den unterschiedlichen Gruppen ausgewählten Wörter bzw. Begriffe kann der Nutzer sein Suchkonzept spezifizieren.

Der Hintergrund dieser IR-Strategie basiert auf der statistischen Wörterhäufigkeit und der Relation zwischen den Wörtern. Die erste Ordnung bzw. direkte Assoziation wird von der Häufigkeit berechnet, in der die Wörter miteinander auftreten. Die Beziehungen der Wörter wie bei Synonymen, Analogien und Antonymen werden durch die zweite Ordnung bzw. indirekte Assoziation hervorgehoben (vgl. [ACKE00]).

Die Clusterstrategie wird dazu verwendet, um die deutliche Beziehung zwischen den Wörtern auf dem Bildschirm zu zeigen und gleichzeitig in der Gruppe zu klassifizieren. Die deutsche SENTRAX funktioniert in der Praxis ziemlich gut. Durch die Mensch-Maschine-Schnittstelle kann man standardisierte Precision- und Recallwerte leider nicht messen.

2.5.1 Design

Weil die Relationsmatrix zwischen den Ausdrücken in der SENTRAX auf dem Korpus aufgebaut ist, kann man den Prozess der SENTRAX in drei Phasen unterteilen.

1. Lernphase
2. Retrievalphase
3. Klassifizierungsphase

In der Lernphase wird die Worthäufigkeit im ganzen Korpus gezählt. Durch den ersten und zweiten Ordnungsprozess wird die Relationsmatrix zwischen den Ausdrücken bzw. den sogenannten Wissensbasen aufgebaut. Die Stoppwörter werden herausgefiltert. Jedes Wort in der Relationsmatrix hat ein Beziehungsgewicht mit den Dokumenten in der Sammlung. In dieser

Situation stehen die aus den indirekten Relationsprozess verbliebenen Wörtern als Index zur Verfügung. Beim Update dürfen die neuen Dokumente zur Wissensbasis hinzugefügt werden.

Zur Retrievalphase wird die Suchanfrage oder ein beliebiges Dokument in die SENTRAX gelegt. Die Suchanfrage wird in kleine Abschnitte bzw. Ausdrücke eingeteilt. Die verwandten Ausdrücke werden aus der Wissensbasis herausgesucht. Mit der Hilfe der SpaCAM und des syntaxbasierten Ähnlichkeitsmaßes werden die Wörter fehlertolerant wiedererkannt.

Zur Klassifizierungsphase wird eine Ähnlichkeitsmatrix aus der indirekten Assoziationsmatrix entsprechend der herausgefilterten Begriffe aufgebaut. Das Clusterverfahren wird momentan verwendet, um den Wortschatz zu erzeugen. In diesem Punkt kommt die Singulärwertzerlegung zum Einsatz, damit die gruppierten Begriffe auf dem zweidimensionalen Raum dargestellt werden können. Die ausführliche Details befinden sich in [ACKE00]. Die Grafikdarstellung in der SENTRAX wird unter der Realität des menschlichen Lernprozesses entworfen. Die gut entworfene Mensch-Maschine-Schnittstelle erhöht nicht nur leistungsfähige Suche (siehe Abschnitt 2.4.4.1), sondern hilft dem Nutzer auch, um die suchende Konstruktion des Korpus zu verstehen. Außerdem kann der Nutzer das Verfahren vom Suchsystem selbst beherrschen, um seinen Wille zum Ziel ohne Ablenkung zu erreichen.

2.5.1.1 Wörterbeziehung

Der Aufbau der Relationen zwischen den Wörtern im Korpus hängt von zwei Verfahren ab, der direkten Assoziation bzw. Assoziations-Hypothese erster Ordnung und der indirekten Assoziation bzw. Assoziations-Hypothese zweiter Ordnung. Die beiden Schritte sind die wichtigen Kernverfahren des ContextMaps, eine Funktion in SENTRAX, um die statistischen assoziierten Beziehungen zwischen den Wörtern zu erhalten.

Da nicht alle Wörter die Bedeutung übertragen, werden die Wörter durch die Stoppwortregel erst gefiltert. Am Anfang wird jedes verbleibende unterschiedliche Wort im Korpus gezählt, um die Häufigkeitstabelle zu erstellen. Die Kookkurrenzhäufigkeit wird auch zusammengerechnet, indem die Kookkurrenzwörter im Kontextfenster mit der Länge $2\delta + 1$ detektiert werden, wobei δ eine Anzahl von Wörtern repräsentiert. Für das Wort „*a*“ und das Wort „*b*“, ist n_a die Häufigkeit des Wortes „*a*“ sowie n_b die Häufigkeit des Wortes „*b*“ im Korpus und

n_{ab} die Kookkurrenzhäufigkeit zwischen den Wörtern „*a*“ und „*b*“, wobei n_{ab} durch einen spezialen Parameter definiert wird (siehe Formel 33; Anhang 7.5).

Anhand der Wörterhäufigkeit und Kookkurrenzhäufigkeit wird die direkte Assoziation berechnet. Die direkte Assoziationsstärke wird durch die Formel 34 (Anhang 7.5) definiert.

Normalerweise wird die „κ“-Konstante ungefähr bei $\frac{2\delta}{n}$ angesetzt. [ACKE00] hat den Wert „κ“ mit $0{,}9 \cdot \frac{2\delta}{n}$ für den Taz-Korpus und sein Modell verwendet.

Nachdem die direkte Assoziationsstärke $ass(a,b)$ für alle beliebige Wörter „*a*“ und „*b*“ im Korpus berechnet worden ist, wird die direkte assoziierte Matrix „**A**“ aufgebaut, wobei $a_{ij} = ass(i, j)$. Danach wird die indirekte Assoziation durch ein Maß für semantische Ähnlichkeit berechnet (siehe Formel 35; Anhang 7.5).

Die indirekte Assoziation bewirkt, dass die Wörter in die angemessene Relation gesetzt werden können, indem die Relation zwischen zwei beliebigen benötigten Wörtern angedeutet wird.

Die ersten n stärker assoziierten Wörter aus der indirekten Assoziation werden durch die Singulärwertzerlegung und die Clustermethode geführt, um die semantische Gruppe in Cluster zu verpacken. Nach dem Cluster-Verfahren und der Singulärwertzerlegung können die semantischen Wortgruppen auf dem Bildschirm angezeigt werden.

2.5.1.2 Menschliche Lernmethode

Wie kann man eine Neuigkeit kennen lernen? Diese herkömmliche Frage stellt sich, wenn das Verfahren zur Beschaffung von Wissen erforscht wird. Das grundlegende Lernen zur Anerkennung, Klassifizierung und Verbindung zwischen alter und neuer Kenntnis basiert auf der Mengenlehre, in der die verschiedenen Eigenschaften des Objektes oder des Musters definiert werden, um die Muster zu gruppieren.

Die übliche einfache Methode zur Gruppenbildung der Objekte ist der Vergleich der Eigenschaften. Im Gegenteil kann man die vertretenden Begriffe jeder Gruppe durch die Eigenschaften identifizieren, ohne eine zusätzliche Erläuterung zu haben, wenn die Objekte eindeutig einer Gruppe zugeordnet wurden. Z.B. kann die Wortmenge {Geschirr, Ofen, Herd, Messer, kochen} den Begriff „Küche“ umschreiben.

Nehmen wir an, wenn man in einen Supermarkt geht, weiß man bestimmt, in welcher Abteilung man nun steht. Anhand der Betrachtung der Umgebung kann man einfach erkennen, wo man im Supermarkt ist und was man in der Umgebung finden kann.

Die Gruppierung ist eine einfache Methode Verstehen zu erleichtern. Zurück zum obengenannten Beispiel. Wenn man eine Party am Abend machen möchte, weiß man, was man kaufen muss und in welcher Abteilung die nötigen Sachen gefunden werden können. Mit dieser Analogie kann man den IR-Prozess vergleichen, indem man das persönliche Suchkonzept auf die allgemeinen und umgebenden Begriffe abbildet. Die gute Einordnung und die Funktion der Konzeptabbildung können die allgemeine Prozesszeit natürlich reduzieren.

2.5.1.3 Konzeptnetz

> The notion has a dual purpose. It comprehends the nature or essence of a subject-matter, and thus represents the true thought of it. At the same time, it refers to the actual realization of that nature or essence, its concrete existence. All fundamental concepts of the Hegelian system are characterized by the same ambiguity. They never denote mere concepts (as in formal logic), but forms or modes of being comprehended by thought.
>
> Herbert Marcuse, *Reason and Revolution* 25.[9]

Das Suchkonzept kann logisch oder psychologisch wahrgenommen werden, abhängig davon, was man begreift, wenn man einen Ausdruck versteht. Einerseits besitzt jedes Konzept eine

[9] Ausgeschnitten von http://www.class.uidaho.edu/mickelsen/texts/Hegel%20Glossary.htm (10.08.05)

eigene Menge unterschiedlicher Eigenschaften. Umgekehrt beschreibt eine Menge der Eigenschaften ein Konzept. Andererseits ist ein Konzept durch Gedanken ergänzend beschrieben. Die Beispiele können auch das Konzept klarmachen. Obwohl die logische Form sehr konkret ist, bauen die meisten Menschen das Konzept mit ihren Gedanken.

> **Concept** - *unit of thought*
>
> The semantic content of a concept can be re-expressed by a combination of other and different concepts, which may vary from one language or culture to another. Concepts exist in the mind as abstract entities which are independent of the terms used to label them.[10]

Für das Informationsretrieval wird das Suchkonzept von den durch die Frage formulierten Suchwörtern konstruiert und von den durch den Anfrageerweiterungsprozess zusätzlichen Attributen ergänzt. Wie im obigen Zitat erwähnt, werden die Attribute bzw. Begriffe bei der SENTRAX durch die ContextMap-Funktion gruppiert und auf dem Bildschirm dargestellt, damit der Nutzer die in einzelnen Gruppen verteilten Attribute wählen kann, um sein Suchkonzept einfach zu bilden.

Während die Attribute von den Gedanken des Nutzers zum Aufbau des Suchkonzepts ausgewählt werden, versteckt sich der logische Vorsatz im Hintergrund. Die formale Begriffanalyse, die mathematische logische konkrete Konzeptsformulierung, wird für die Anfrageerweiterung und das relevante Feedback im Informationsretrieval zunutze gemacht (vgl. [GRWE02] und [GRWE04]). Grootjen hat das logische Konzeptnetz von der formalen Begriffanalyse aufgebaut. Seine Konzeptsknoten sind das unterschiedliche Paar von einer Dokuments- und Attributsmenge. Obwohl unser Ansatz nicht auf der formalen Begriffanalyse basiert, kann sie unseres Experiment aber erklären.

Also, bei unserem Ansatz wird das Konzeptnetz als verteiltes Grundkonzept durch die Funktion von ContextMap betrachtet. Unser Konzeptnetz wird durch die verbundenen Eigenschaf-

[10] Ausgeschnitten von www.willpowerinfo.co.uk/glossary.htm (10.08.05)

ten verfasst. Die gruppierten Eigenschaften bzw. Begriffe stammen aus den Wörtern im Korpus und stellen deren Beziehungen dar. Anhand der Auswahl vom Nutzer wird ein nutzerbezogenes Suchkonzept verwirklicht. Durch die Auswahl kann ein psychologisches Konzept exakter beschrieben werden.

2.5.1.4 Verbindung zur realen Suche

Beim IR-Prozess benötigt man die mit dem gedanklichen Suchkonzept bestimmten Dateien, die in einem Korpus oder in mehreren Korpora zusammengelegt werden. Die Indexe werden von ausgewählten Wörtern aus den Dokumenten des Korpus gesammelt. Um auf die Dokumente zuzugreifen, wird das Ähnlichkeitsmaß zwischen Anfrage und Index durch mathematische oder statistische Methoden berechnet.

Damit die Kommunikation zwischen dem Computer und dem Nutzer zum alltäglichen Leben ähnlich ist, verwendet die SENTRAX die wörtliche Bedienoberfläche, mit der der Nutzer die seiner Anfrage entsprechende Wörterbeziehung klar erkennen kann. Anhand dieser Methode darf der Nutzer die in der Beziehung stehenden Wörter aus verschiedenen Begriffsgruppen wählen, um die Eigenschaft der benötigten Dokumente zu kombinieren, wie die Auswahl der Materialien im Baumarkt, um den Schrank oder Tisch wie in seinem Entwurf aufzubauen.

Die Zusammenbindung zwischen den Theorien im Hintergrund und der Anfrage mit einer solchen Mensch-Maschine-Schnittstelle hilft dem Nutzer nicht nur, um die Relationen zwischen seiner Anfrage und den entsprechenden Indexen zu verstehen, sondern auch, um die beinhalteten Themen im Korpus entsprechend seiner Anfrage zu entdecken. Außerdem kann so die Zugriffzeit verringert werden. Das bedeutet, dass der Nutzer an die mathematischen oder statistischen Methoden, die im Hintergrund stehen, nicht gewöhnt sein muss, weil die Mensch-Maschine-Schnittstelle die Vermittlung der wörtlichen Beziehungen in den Dokumenten darstellt.

Insbesondere für den Korpus, der von gleichartigen, vergleichbaren und verwandten Themen handelt, werden die Wörter zusammengelegt. Dieser typische Korpus kann über das Konzeptnetz herausgefunden werden, um die benötigten Dokumente wieder auffinden zu können.

Die mehrdimensionale Begriffswolke wird als eine Art semantischer Index in einer Datenbank abgelegt. In einem Beispiel sieht die Begriffswolke entsprechend der Suchanfrage „VW“ wie folgt aus: (siehe Abbildung 2)

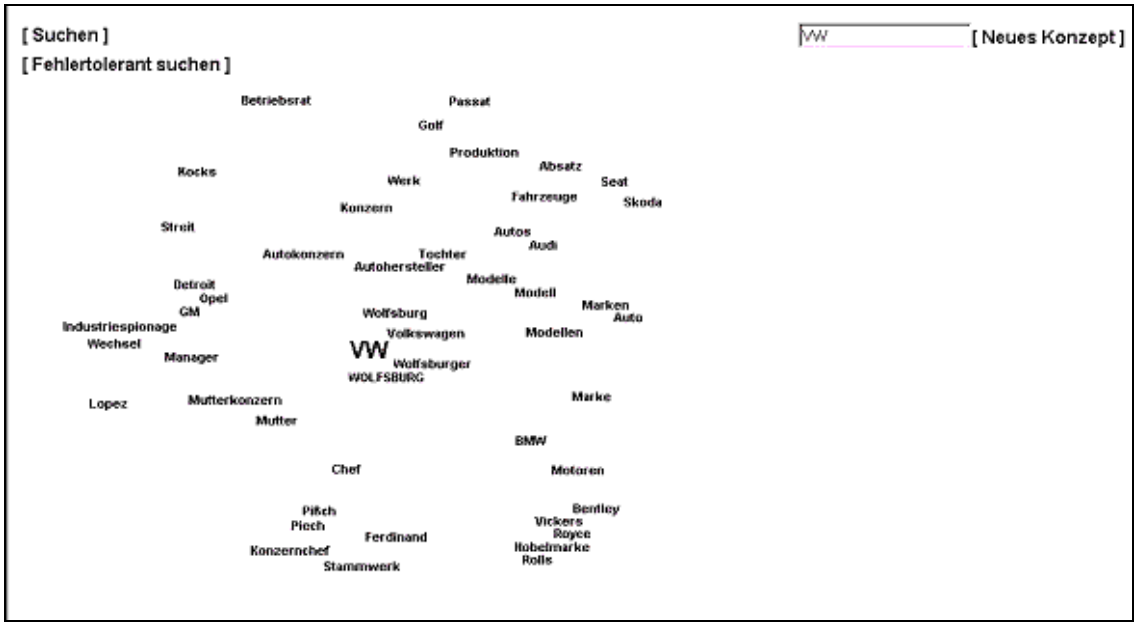


Abbildung 2 Die erzeugte Begriffswolke unter der Suchanfrage „VW“ in einer frühen Version der SENTRAX

Man kann sehen, dass die Relationen semantisch locker geteilt werden, z.B. Automarke (Seat, Skoda, Audi), Geschehen (Lopez, Industriespionage, Wechsel, Manager), Modell (Golf, Passat), Konzern (Mutterkonzern, Tochter, Autohersteller), Motor (BMW, Bentley). Solche Extraktionen sind also kein explizites Cluster.

2.5.2 Funktionen der SENTRAX

Vier wichtige Funktionen befinden sich in der SENTRAX. In den folgenden Absätzen lassen sich die Wirkungsweisen der vier Funktionen kurz beschreiben.

2.5.2.1 LexicoMap-Funktion

Drei Möglichkeiten der Wortfehler sind Tippfehler in der Anfrage, Tippfehler bzw. OCR-Fehler in den Dokumenten und vielfältige Schreibweise (z.B. alte bzw. neue Schreibweise und tolerierte Schreibvarianten). Diese Probleme können beim Informationsretrieval immer

passieren. Die SENTRAX bietet dem Nutzer die orthographische Funktion LexicoMap, um andere Schreibweise bzw. auch die Fehler und Richtigkeit aufgedeckt zu bekommen. Diese Fehler können von den invertierten Listen nicht entdeckt werden. Hingegen werden die Schreibvarianten und Tippfehler allerdings durch die LexicoMap dargelegt [BENT06]. Als Wichtigstes wird das Kompositum auch erkannt, falls es im Korpus enthalten ist.

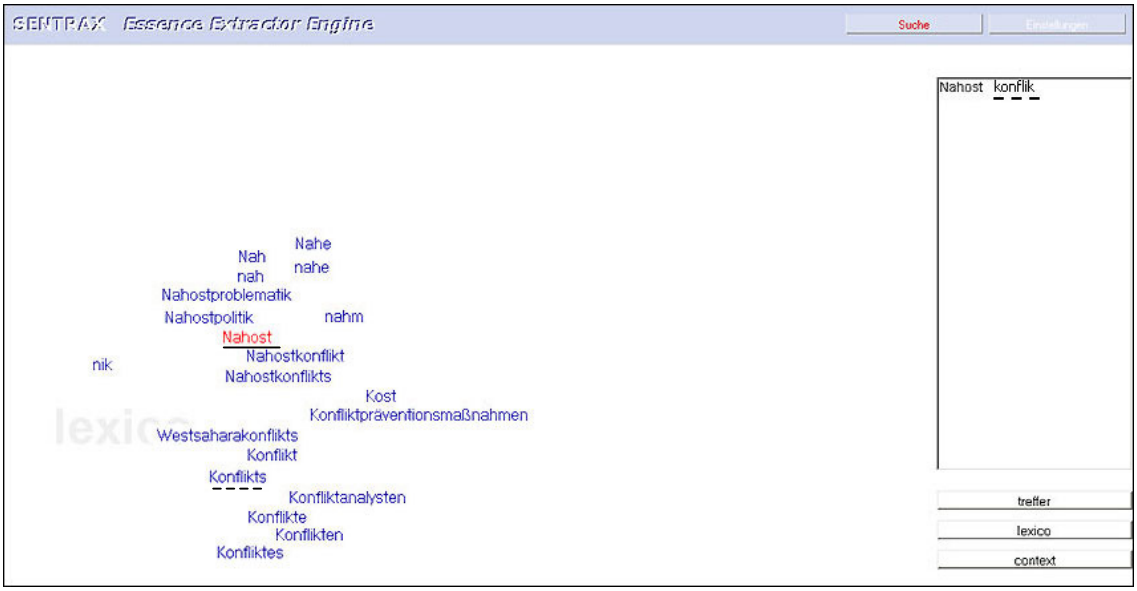


Abbildung 3 LexicoMap : Mit Eingabe – Nahost und Konflik (Tippfehler) – bringt die LexicoMap sowohl andere ähnliche Schreibvarianten als auch das Kompositum „Nahostkonflikt“ zum Vorschein.

2.5.2.2 ContextMap-Funktion

Die direkten und indirekten Assoziationen von Wörtern stammen aus der Kookkurrenzanalyse des ganzen Korpus. Die Zusammenhänge der Wörter werden in Gruppen durch das Clusterverfahren und die Singulärwertzerlegung verteilt. Weil die Wörter aus dem Korpus genommen werden, stehen nur die Wörter, die im Korpus gefunden werden können, in der ContextMap. Obwohl dies die Thesauri für den Nutzer beschränkt, kann man sicher sein, dass jedes aufgetauchte Wort zu irgendeinem Dokument gehört und die Wörter dem Nutzer helfen können, um den groben Zusammenhang zu bilden. Mit Hilfe von unscharfen assoziativen Wörtern durch die zweidimensional dargestellte Mensch-Maschine-Schnittstelle realisiert die ContextMap-Funktion eine auffällige leistungsfähige Suche. Die ContextMap-Funktion kann dem Nutzer helfen, um den Sinn der Bedeutungs- und Realbezüge zu erkennen. Die auf dem

Bildschirm auftauchenden Wörter, die räumlich angeordnet werden, werden „Begriffswolke“ genannt. Ob der Nutzer nun Novize oder Experte ist, es hilft ihm die ContextMap-Funktion seine Idee durch vorkommende Wörter zu erweitern. Anhand der kartographischen Vernetzung der Informationsbestände kann man entweder einen beliebigen Begriff innerhalb einer Begriffswolke anklicken oder einen unnötigen Begriff ausklicken, um eine neue Begriffswolke zu generieren. Falls kein Erweiterungswort auftaucht, kann der Nutzer neue Attribute natürlich ins Anfragefenster direkt eingeben. Falls die neuen Attribute nicht im Korpus gefunden werden, werden sie automatisch vom Anfragefenster gelöscht. In einem solchen Fall bleibt die Begriffswolke unverändert.

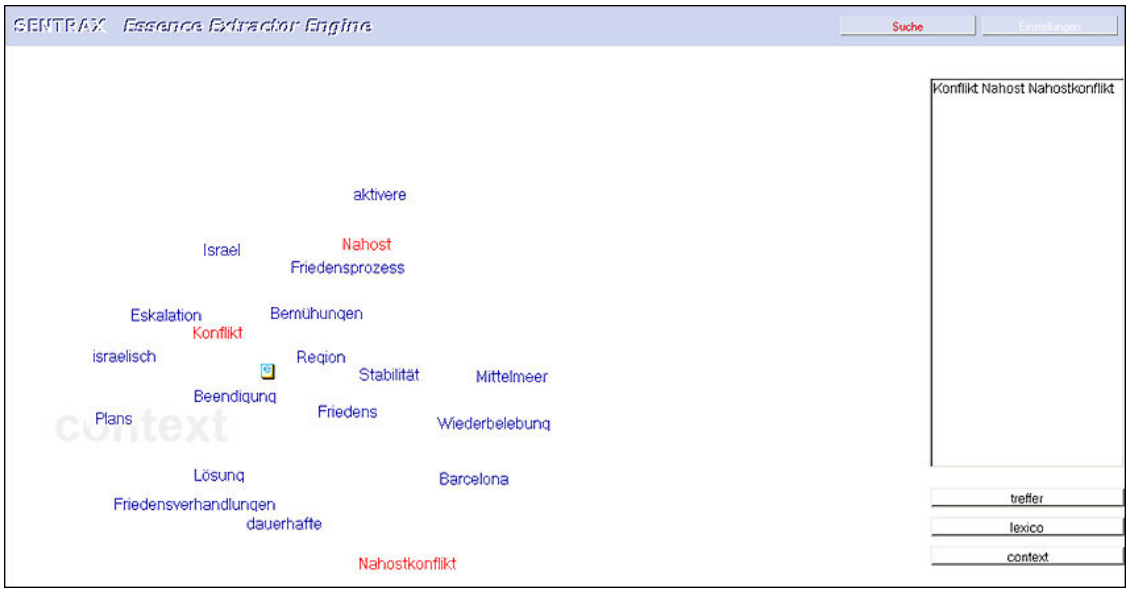


Abbildung 4 ContextMap : Nach der Auswahl der Attribute „Konflikt“, „Nahost“ und „Nahostkonflikt“ bildet die ContextMap die häufigen zusammentreffenden Attribute in die Gruppe. Man kann sich vorstellen, von welchen Themen die Dokumente mittels der vorkommenden Attributen handeln können.

Die Begriffswolke bzw. die Gruppe der Wörter kann dem Nutzer helfen, um zugehörige Begriffe abzufragen. Diese sind nicht festgelegt wie in der Taxonomie oder im IS-A-Thresaurus, sondern veränderlich entsprechend den Hauptbegriffen. Die Wörter bzw. Begriffe, die statistisch eng verwandt sind, werden nah beieinander platziert. Je schwächer der Zusammenhang, desto entfernter sind die Begriffe. Durch das semantisch visualisierte Verknüpfungsnetz kann der Lernprozess zwischen dem Nutzer und der Begriffwolke das Verständnis vom Korpus einfach bilden und zur effizienten Entscheidung für die richtige Wissensdomäne führen. Au-

ßerdem kann man in der Begriffwolke bedeutungsverwandte Begriffe auffinden, die bei invertierten Listen gar nicht entdeckt werden können [BENT06].

2.5.2.3 TrefferDoc-Funktion

Die TrefferDoc-Funktion liefert die Dokumente entsprechend der Suchwörter in der Rangliste mit einer Prozentangabe des Zusammenhangs zwischen den Dokumenten und den Suchwörtern. Ein beliebiges IR-Modell kann hier als Hintergrund verwendet werden. Für die aktuelle SENTRAX wird das boolesche Modell verwendet.

2.5.2.4 SimilarDoc-Funktion

Die letzte Funktion, SimilarDoc-Funktion, ist ein Feedback für den Nutzer, der die ähnlichen Dokumente mit dem ausgewählten Dokument aus der Trefferliste ermitteln will. Diese Funktion befindet sich am Ende des Dokumentbeispiels in der Trefferliste. Diese Funktion prüft die Ähnlichkeit der Indexe zwischen den Dokumenten.

2.5.3 SENTRAX im Vergleich zum klassischen Modell

In den Abschnitt 2.5.1 und Abschnitt 2.5.2 erkennt man bereits, welche Strategie in der SENTRAX verwendet wird und wie sie funktioniert. Als IR-Anwendung kann die SENTRAX mit anderem klassischen Modell verglichen werden. Einige Merkmale sind die folgenden:

- Bei der klassischen Methode wird die Anfrage meist auf den Indexausdrücken durch die Ähnlichkeitsfunktion direkt abgebildet. Im Gegenteil dazu werden die verwandten Suchkonzepte von der Anfrage bei der SENTRAX zunächst erzeugt, damit der Nutzer seine neue passende Anfrage auswählen kann. Die Anfrage wird mittels der Auswahl aus dem Konzeptnetz bestimmt und gleichzeitig erweitert.
- Die von direkter und indirekter Assoziation gestaltete Beziehung der Wörter wird von der Nutzung der Wörter in einer Umgebung erschaffen. Solche Beziehungen sind von dem Kontext abhängig. Der Zusammenhang der Wörter, der auf der Ebene gezeigt wird, könnte einem Nutzer helfen, um die Charakteristik der im Korpus gelegter Texte

grob aufzuspüren. Die Auswahl von aufeinander bezogenen Wörtern kann ein individuelles Suchkonzept vertreten. Sie erlaubt auf die wahrscheinlich relevanten Dokumente einfach zugreifen. Im Vergleich zu der SENTRAX lässt sich die Anfrage bei klassischem Modell durch die Erweiterungstechnik durchführen. Die verwandten Wörter, ob sie von Kontextanalyse oder semantischem Netz oder anderer Methode erhalten werden, werden mit der originalen Anfrage vereint, um eine Chance zum Zugriff zu den Dokumenten mit verwandten Wörtern zu haben. Der Erweiterungsprozess versteckt sich aber im Hintergrund. Der Nutzer hat keine Chance, den Prozess zu erkennen oder seine Anfrage nach der Erweiterung zu sehen. Deshalb werden die unrelevanten Dokumente in der Rangliste vielleicht häufiger auftauchen, obwohl der Nutzer eine solche Erweiterung nicht braucht.

- Die Wörter, die auf der zweidimensionalen Ebene gezeigt werden, werden von der ermittelten Kookkurrenzhäufigkeit im Erkennungsfenster fester Größe abgeleitet und durch die Clustermethode, nämlich die Ward-Methode, klassifiziert. Der Vorteil davon ist, dass die Relation, beispielweise Synonyme, Antonyme usw., darstellbar macht [ACKE00]. Das hilft dem Nutzer, seine Anfrage zu erweitern. Gleichzeitig hilft die Verteilung durch die Clustermethode dem Nutzer die Relationen der Wörter zu verstehen und die versteckten Konzepte hinter den Wortgruppen einfach zu vorstellen.
- Statt N-Gram allein wird die SpaCAM-Technologie verwandt, um die Fehler der Schreibweise automatisch zu erkennen. Ein großer Vorteil dabei ist, dass Tippfehler an beliebigen Positionen wiedererkannt werden können.
- Nicht nur Varianten der Schreibweise können durch das im Korpus enthaltene deutsche Kompositum durch die LexicoMap-Funktion aufgefunden werden. Wenn man beispielweise die Suchwörter „Nahost“ und „Konflikt“ eingibt, findet die LexicoMap-Funktion das Kompositum „Nahostkonflikt“. Die herkömmliche Präfix- und Suffixerkennung können ein solches Kompositum nicht angegeben.
- Das Lernverfahren während des Suchprozesses bewirkt, dass der Nutzer die Relationen zwischen den verwendeten Suchwörtern und ihren verwandten Attributen verstehend verfolgen und sich die im Korpus enthaltene Informationen vorstellen kann. Aber die klassischen Modelle können kein solches Lernverfahren anbieten.

- Obwohl im Vergleich zur Bildung der Cluster des Dokumentes die Cluster der Wörter während der Laufzeit Extrazeit kosten, ist das Laufzeitverhalten nicht so kritisch, weil die Wörter in der Matrixform gespeichert werden und die Berechnung des Matrixteils nicht sehr lange dauert. Bei dem Dokumentenclusterverfahren wird der Zugriff eventuell von einer Anfrage in ein ungeeignetes Dokumentencluster umgeleitet. In diesem Fall würde das Retrieval möglicherweise viele unrelevante Dokumente liefern.
- Singulärwertzerlegung ist ein nützliches Werkzeug, um die Konzeptsdimensionen im LSI-Modell zu verschmelzen. Bei der SENTRAX wird Singulärwertzerlegung auch verwendet, um in ähnlichen Wortcluster zu zerlegen. Die Cluster des Wortes, die nicht zu große Unterschiede haben, können in einem Cluster verpackt werden.
- Für die Komplexität der Singulärwertzerlegung gilt:

  i. Allgemeine Formel für die Komplexität der exakten Singulärwertzerlegung: $O\left(m^2n^3 + m^3n^2\right)$ für die Matrix mit der Größe $m \times n$.

  ii. Für das LSI-Modell (vgl. Abschnitt 2.2.6): $O\left(N^2s^3\right)$ mit N = Anzahl aller Dokumente in der Sammlung und s = gewünschte Dimension des Semantikraums.

  iii. Für die SENTRAX mit der exakten Singulärwertzerlegung: $O\left(d^2t^3 + d^3t^2\right)$ mit t = gewünschte Anzahl der anzuzeigenden Wörter und $d = 2$, weil die Wörter auf eine Ebene projiziert werden müssen.

  Weil $t << N$ und $d = 2$ gilt, ist der benötigte Laufzeitaufwand bei der SENTRAX deutlich geringer als bei dem LSI-Modell.

3 KROSSLINGUALE SUCHE

3.1 Grundlage

Das Wirtschaftswachstum hat das Interesse der Firmen und Organisationen, ihr Engagement im Ausland auszubauen, deutlich gesteigert. Die Menge des Verbrauchs an fremden Informationen steigt somit in der gesamten Welt. Daher wird die internationale Zusammenarbeit immer wichtiger. Außerdem zwingt das Internetwachstum den elektronischen Handel, die Ausbildung, die Forschung usw. zur ständigen Erweiterung. Ein Teil der Entwicklung der Digitalbibliotheken wird auch von der krosslingualen Informationsrückgewinnung unterstützt. Daraus resultiert das Bedürfnis an multilingualen Informationen in allen Bereichen, sowohl im öffentlichen als auch m privaten Sektor. Mit diesem Bedürfnis haben sich drei große Institutionen (TREC, CLEF, NTCIR[11]) beschäftigt. Die Entwicklung der krosslingualen Informationsrückgewinnung findet sich manchmal unter einer IR-Institution wie z.B. SIGIR (Special Interest Group on Information Retrieval). Die Verbindungen zwischen dem krosslingualen Informationsretrieval (Abk. CLIR) und dem Informationsretrieval (Abk. IR) sind sehr eng, weil das CLIR den Rahmen von IR auf mehrere Sprachen erweitert.

3.1.1 Grundidee

Das Problem des Zugriffes auf eine multilinguale Datenbank kann als Erweiterung des generellen Informationsretrievals durch die Paraphrase betrachtet werden [FLUH04]. Die zugrundliegende Idee stammt aus der Verbindung von zwei oder mehr IR-Systemen. Es gibt sehr große Informationsmengen, die in unterschiedlichen Sprachen geschrieben wurden.

---

[11] TREC Text REtrieval Conference, CLEF Cross-Language Evaluation Forum, NTCIR NACSIS Test Collection for IR Systems.

Das Verlangen nach krosslingualer Information entsteht beispielsweise daraus, dass:

man mehrsprachige Texte vergleichen will.

man die Richtigkeit des einheimischen oder fremden Textes prüfen will.

man ähnlich thematische Dokumente in einer anderen Sprache braucht.

man die einheimischen Texte anhand eines fremdsprachlichen Stichworts suchen will.

man mehr Information in der anderen Sprache braucht, aber man kennt den fremden Begriff nicht gut genug.

Durch die Übertragung der Anfrage von einer Sprache in eine andere versucht man, die geeignete wörtliche Umsetzung zu ermitteln, damit die passenden Dokumente in den unterschiedlichen Sprachen gefunden werden können. Diese konventionelle Idee stammt aus der Verbindung zwischen zwei IR-Systemen über die wörtliche semantische Übertragungsbrücke.

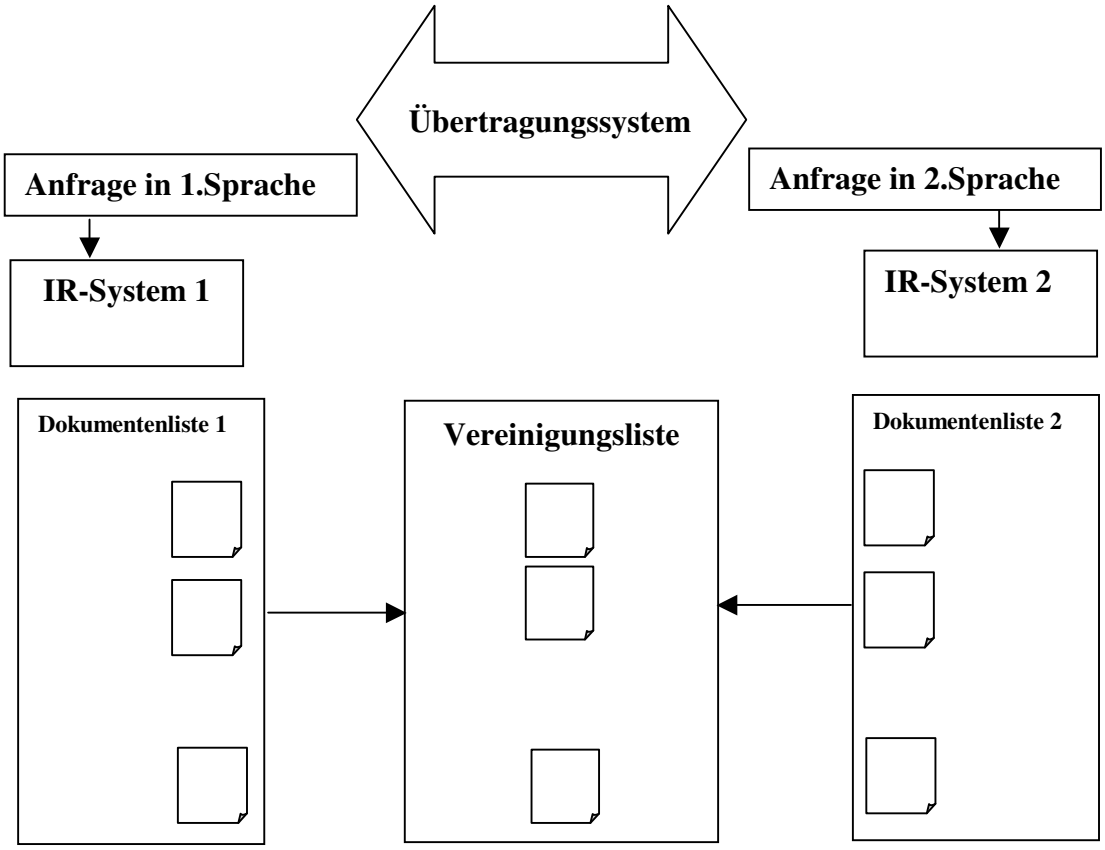


Abbildung 5 Einfaches konventionelles bilinguales Suchmodell.

Zwei wichtige Punkte kann man in der obigen Abbildung sehen. Der erste Punkt ist das Übertragungssystem bzw. die Übertragungsmethode. Der zweite Punkt ist das Vergleichmaß der endlichen zusammengeführten Liste. Die Erklärung der beiden Punkte wird im Abschnitt 3.1.2 und 3.1.3 beschrieben.

Zwei Faktoren, die die Leistungsfähigkeit des CLIR-Systems beeinflussen könnten, sind:

- Die Mehrdeutigkeit sowohl bei den monolingualen Systemen als auch bei den krosslingualen Systemen. Besonders bei den krosslingualen Systemen, die nach der Übertragungsmethode arbeiten, entsteht häufig das Problem der mehrdeutigen Übersetzung.

- Mangelhafte bzw. beschränkte Ressourcen bilden zum Beispiel das bilinguale maschinenlesbare Wörterbuch, der parallele Korpus, die Entwicklung des IR-Systems für einige Sprache usw.

Die Übertragungsmethode kann mit einer einzelnen Brücke zwischen zwei Sprachen, mit der zwei IR-Systeme sich verbinden können, verglichen werden. Der wichtigste Anteil in einem krosslingualen IR-System ist die Übersetzung der Anfrage. Insofern legen die meisten Forschungen ihr Interesse auf die Reduktion der Mehrdeutigkeiten bei der Übersetzung. Probleme entstehen hier nicht nur durch Synonyme, die Polysemie und die Homonymie, sondern auch durch fehlerhaft übersetzte Wörter, Mehrwortgruppen und nicht-gefundene Wörter. Für die CLIR hängt die Mehrdeutigkeit auch bei der Erweiterung der Anfrage davon ab, wie gut die Kenntnis des Suchers von der fremden Sprache ist, weil er die übersetzten Wörter als anfrageerweiterte Terme wählt [BNX98].

3.1.2 Übertragungsmethode

Bei der Übertragungsmethode gibt es verschiedene Strategien, um die semantische Bedeutung der Anfrage in einer anderen Sprache weiterzuleiten. Die zusammengeführte Liste wird aus mindesten zwei verschiedenen Dokumenten zusammengestellt. Das Vergleichmaß bewertet die Leistung des Systems. Beim CLEF Projekt wird die normale Precision- und Recallbewertung verwendet [PETE01]. Wenn das Suchverfahren von beiden IR-Systemen gleich ist, wird angenommen, dass die Leistung nur von der Übertragungsmethode abhängt.

Die Übersetzung bzw. Übertragung ist sehr wichtig für das krosslinguale IR-System. Der Hauptschlüssel der Leistung hängt davon ab, wie gut bzw. genau die übertragene Anfrage ist. Es gibt hierbei drei Fehlerquellen. Die erste ist die Mehrdeutigkeit von „Wort-Wort-Übersetzung“, zahlreichen abgeleiteten Wörtern und verwandten Bedeutungen. Die zweite ist der semantische Verlust wegen Übertragung der Wortgruppe der Anfrage durch die Wort-Wort-Übersetzung. Die dritte ergibt sich dann, wenn für einige Wörter keine Übersetzungen gefunden werden können.

3.1.2.1 Wörterbuchbasis

Das Wörterbuch basierte Vorgehen besteht darin, dass die Anfrage von einer Sprache in eine andere Sprache durch ein maschinelles lesbares Wörterbuch übersetzt wird. Aufgrund des Bedarfs an der fremden Informationen ist die Größe des elektronischen bilingualen Wörterbuchs beständig gestiegen. Trotzdem es nicht alle Domänen komplett umfasst, ist es einfach und praktisch anzuwenden. Die übersetzte Anfrage kann dabei unter Umständen im Korpus enthalten sein. Manche verwenden nur ein bilinguales elektronisches Wörterbuch, aber manche nutzen die maschinelle Übersetzung, da sie nicht nur die Anfrage übertragen, sondern auch das Zieldokument übersetzen wollen.

Bei der Wörterbuchbasis entsteht ein großes Problem, wenn das allgemeine Wörterbuch für einen Fachgebietskorpus verwendet wird. Weil das Wörterbuch nur einen allgemeinen Wortschatz hat, findet keine oder nur eine unzulässige Übersetzung bei Fachbegriffen statt. Wie bei den kontrollierten Wörtern (siehe Abschnitt 3.1.2.2), gibt es auch bei der Wörterbuchbasis das Problem der Mehrdeutigkeit, weil vielfältige Übersetzungsmöglichkeiten für ein Wort gegeben sind und eine Wort-Wort-Übersetzung nicht präzise genug ist [GNXZZH98]. Außerdem ist die Übersetzung der Mehrwortgruppe durch das maschinelle lesbare Wörterbuch sehr beschränkt [JARU01]. Obwohl die Verbindung zwischen zwei IR-Systemen durch das elektronisch lesbare Wörterbuch die Leistung des krosslingualen IR-Systems 40-60% weniger als bei dem Monolingualsystem beeinflusst [BALL00], versuchen viele Forschergruppen zusätzliche Strategien zu integrieren, z.B. mit Korpusbasis, mit statistischer Mehrwortgruppenerkennung [GNXZZH98], weil elektronisch lesbare Wörterbücher heute bequem für die verschiedenen Sprachen gefunden werden können. Solche werden von Zeit zu Zeit erneuert und verbessert.

3.1.2.2 Kontrollierte Wörter

Bei dieser Methode werden die wichtigen konzeptuellen Begriffe der verschiedenen IR-Systeme als kontrollierte Wörter festgelegt. Die wichtigen konzeptuellen Begriffe erhält man aus dem Deskriptor des Dokumentes. Der Deskriptor bzw. der Begriff bezieht sich

auf ein Schlüsselwort, wie z.B. der Verfasser oder ausgewählte Ausdrücke. Alle Begriffe in den verschiedenen Sprachen werden zu allgemeinen konzeptuellen Begriffen übersetzt, die von der Sprache unabhängig sind. Die semantische Relation, beispielsweise Synonyme, relevante Ausdrücke, enge oder breite Ausdrücke, kann verwendet werden, um gute Begriffe zu bekommen [FLUH04].

Die Methode der „kontrollierten Wörter“ für das krosslinguale Informationsretrieval ist sehr erfolgreich in kommerziellen und amtlichen Anwendungen[OARD97b]. Oard hat jedoch einen zu großen Korpus in der gleichen Veröffentlichung als nachteilig erwähnt. Ein anderer Nachteil wird in [FLUH04] beschrieben. Es geht darum, dass einzelne Ausdrücke von einer Sprache in unterschiedliche Ausdrücke in der anderen Sprache gebracht werden müssen. Zwei Beispiele werden hier genannt: Das Wort „Katze“ bedeutet auf Englisch „cat (*Tierart*)“ oder „trolley (*Zettelmaschine*)“, das englische Wort „body“ kann aber auch die Bedeutung „die Leiche“, „der Körper“, „der Aufbau“, „Gremium (Körperschaft)“, „Gehäuse“ oder „Wagenkasten (Eisenbahn)“ usw.[12] besitzen. Dieses Problem wird „Polysemie“ genannt (vgl. [JUMA00]).

Die semantische Domänenmarkierung, beispielweise Tier, Maschine, Anatomie usw., könnte die Übersetzung abhängig vom Nutzungszweck des multilingualen Thesaurus beeinflussen und hilft vielleicht dabei den multilingualen Thesaurus zu erzeugen.

3.1.2.3 Korpusbasis

Anhand der Korpusbasis kann man das Problem der verallgemeinerten Übersetzung aufgrund der Wörterbuchbasis beseitigen. Weil ein bilinguales Wörterbuch verwendet wird und es unpraktisch ist, dies selbst zu erstellen, analysiert die Korpusbasis eine große Textsammlung [OARD97b], indem die Wörter von parallelen oder vergleichbaren Textsammlungen automatisch extrahiert und ihre Relationen bewertet werden.

[12] Die Übersetzung von http://dict.leo.org/

[OARD97b] diskutiert zwei Faktoren der Beschränkung. Der Mangel an parallelen Korpora stellt eine wesentliche Einschränkung dar. Deshalb müssen häufig Textsammlungen zunächst manuell übersetzt werden, um parallele Korpora zu erzeugen. Es wird angenommen, dass man existierende vergleichbare Texte automatisch erkennen und die Grenze der Korpusbasis mit der vergleichbaren Textsammlung erweitern kann. Eine andere Beschränkung ist abhängig davon, wie gut die extrahierende Methode ist. Die bisher beste Methode ist wahrscheinlich die Integration der sprachwissenschaftlichen und statistischen Methoden.

3.1.2.4 Vektorraummodell von Salton

Dieses Modell stammt von Salton [FLUH04]. Die Dokumente werden in einem n-dimensionalen linearen Raum betrachtet. Die Größe des Raums ergibt sich aus der Anzahl der unterschiedlichen Wörter. Es wird behauptet, dass der Unterraum der relevanten Dokumente in einer Sprache in Relation steht mit dem Unterraum der übersetzten Dokumente in der anderen Sprache. Wenn also die Anfrage übersetzt und in das eigene IR-System durchgeführt wird, könnten die Unterräume der übersetzten Anfrage die relevanten Dokumente bezüglich der Zielsprache zurückliefern. Die Ähnlichkeit zwischen dem Dokument und der übersetzten Anfrage kann mit dem Skalarprodukt gemessen werden.

Diese Methode wird mit dem Latent-Semantic-Indexing (LSI) bei Lanndauer und Littman kombiniert. [LALI90] nutzt 900 Abschnitte von 2482 englisch-französisch bilingualen Korpora, um die Matrizen zu trainieren. Die 1582 französischen Dokumente werden in die Matrizen eingefügt. Der Rest der englischen Dokumente wird als Anfrage eingegeben. Als Ergebnis wurde zu 92% korrekt den analogen französischen Dokumenten zugeordnet. In [DLLL97] wird der multilinguale semantische Raum durch die LSI-Methode abgearbeitet, um die französisch-englische Textsammlung zu testen. Das Ergebnis wird auch mit der maschinellen Übersetzung verglichen. Die parallelen Texte werden benutzt, um das krosslinguale LSI-IR-System zu strukturieren. Weil die Ausdrücke und Dokumente auf den multilingualen semantischen Raum gelegt werden, können die Dokumente entsprechend ihrer nicht übersetzten Anfrage in beliebiger Sprache

zurückgeholt werden. Um das entsprechende Dokument zu finden, ergibt sich mit über 98% das richtige Paar auf dem ersten Rangplatz. Im Vergleich zum herkömmlichen Vektormodell ohne LSI-Methode wird die Wahrscheinlichkeit ca. 100% gesteigert, dass sich das richtige Paar auf dem ersten Rangplatz befindet. Außerdem ergibt sich für die Leistung der Ermittlung des richtigen Paares auf dem ersten Rangplatz zwischen MT-LSI, maschineller Übersetzung mit den zwei LSI-IR-Systemen und dem krosslingualen LSI-IR-System kein Unterschied.

3.1.2.5 Maschinelle Übersetzung

Die maschinelle Übersetzung ist ein bekanntes Werkzeug, um ein Dokument automatisch zu übersetzen. Das Vermögen der maschinellen Übersetzung wird für verschiedene Textformate entwickelt, z.B. Textdatei, Pdf-Datei und Website. Babel-Fish-Translator[13] ist beispielweise eine Website zur kostenlosen Übersetzung mit Wortanzahlbegrenzung. Es stehen auch viele kommerzielle Übersetzungsanwendungen in verschiedenen Sprachpaarungen zur Verfügung. Aufgrund der Fähigkeit und der fortlaufenden Entwicklung von vielen Anbietern ist die maschinelle Übersetzung eine Möglichkeit, um das krosslinguale Informationsretrieval durchzuführen.

Statt eine alleinige normale maschinelle Übersetzung zu verwenden, wurde die Kombination von einer maschinellen Übersetzung und der IR-Methode vom European ESPRIT Consortium, EMIR[14], entwickelt. Das System basiert auf drei hauptsächlichen Werkzeugen [FLUH04]:

- Sprachwissenschaftlicher Prozessor – dadurch werden die Morphologie- und die Syntaxanalyse im Text ausgeführt.

[13] http://world.altavista.com

[14] European Multilingual Information Retrieval

- Statistisches Modell – dadurch wird die Verbindung zwischen Anfrage und Dokument gewichtet.
- Monolinguales oder multilinguales Reformationssystem – dabei handelt es sich darum, die originale natürliche Anfrage in ein analoges Konzept der anderen Sprache zu übertragen.

Die Texte werden zuerst durch den sprachwissenschaftlichen Prozessor durchgeführt, um die Einzelwörter und die Mehrwortgruppen zu erkennen. Das Gewicht wird für alle normalisierten Wörter mit Hilfe des statistischen Modells berechnet. Die Anfrage wird durch den gleichen sprachwissenschaftlichen Prozessor bearbeitet. Danach wird das Ergebnis in den Reformationsprozess geschickt, um den neuen Ausdruck zu ermitteln. Bei [FLUH04] wird noch erklärt, dass der wesentliche Unterschied zwischen dieser Methode und der maschinellen Übersetzung nur die Übersetzungsmethode ist, da bei dieser Methode die Mehrwortanfrage zunächst unterbrochen wird und dann jedes Wort übersetzt wird.

Weil jedes Wort der Mehrwortgruppe übersetzt wird, tauchen mehrdeutige Wörter auf. Anhand dem Vergleich des Zusammentreffens von Wörtern bei anderen Konzepten wird die implizite semantische Information herausgefunden, um die richtige Übersetzung zu finden. Diese Idee wird auch bei [GNXZZH98] verwendet.

Die maschinelle Übersetzung ist für ein einzelnes Dokument geeignet. Dagegen ist sie für alle Dokumente in der Sammlung unpraktisch, während lediglich die Anfrage bei einem krosslingualen IR-System übersetzt wird. Obwohl es eine maschinelle Übersetzung für die Ausgangsprache gäbe, braucht eine maschinelle Übersetzung oft einen längeren Kontext als die übliche Anfrage, um die Übersetzung originalgetreu zu ermöglichen [BALL00].

3.1.2.6 Umsetzungssprache

In dieser Methode gibt es zwei verschiedene Ideen. Die erste Idee stammt aus der Methode der „kontrollierten Wörter“. Die Wörter mit Hilfe der Indexierung automatisch so

zu reduzieren, dass das semantische Konzept noch erhalten bleibt. Die ausgewählten Konzepte werden mit Ausdrücken verknüpft, die im gesamten Text zu finden sind. Die Verbindung zwischen dem Konzept und dem Text kann entweder durch einen manuellen oder einen automatischen Lernprozess erstellt werden. Die Verbindung zwischen dem Konzept und dem fremden Text wird zuerst hergestellt. Aus diesem Grund kann man die Anfrage in der gleichen Sprache wie das Konzept in der multilingualen Datenbank suchen lassen.

Die zweite Idee unterscheidet sich unwesentlich. Wegen des Fehlens eines bilingualen Wörterbuches wird eine bekannte Sprache als Umsetzungssprache, z.B. englisch, festgehalten. Mit der sprachlichen Brücke kann die deutsche Anfrage beispielweise zunächst ins Englische und dann vom Englischen ins Italienische übersetzt werden. In diesem Fall wird die englische Sprache als Umsetzungssprache benannt. Ein paar Beispiele befinden sich im Eurospider Retrieval System [BKSK99] oder Cross-Language Retrieval via Transtive Translation von Ballesteros [BALL00].

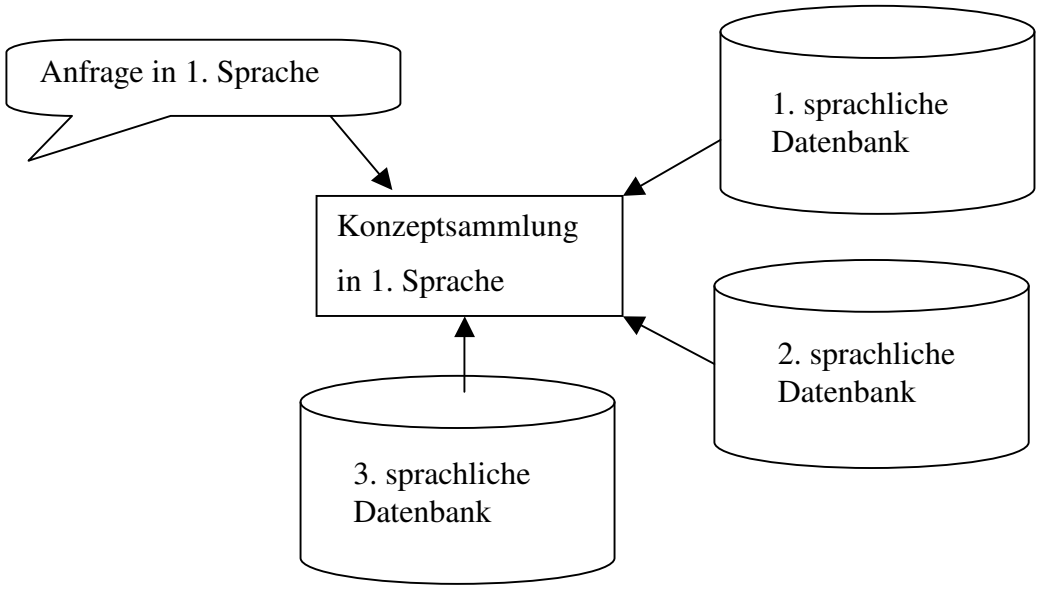


Abbildung 6 Die Struktur der Umsetzungssprache.

Außer den obigen Methoden gibt es noch die Kombination mehrerer Übertragungsmethoden, wie zum Beispiel eine Kombination des bilingualen Wörterbuches mit der Korpusbasis. Die Notwendigkeit zu Kombinationen ergibt sich aus dem Mangel an Fachgebietswörtern und der Mehrwortgruppe in einem allgemein bilingualen Wörterbuch.

[GNXZZH98] haben das bilinguale maschinenlesbare Wörterbuch mit der Korpusbasis zusammen verwendet, um die Mehrwortgruppe der anderen Sprache anstatt ohne Wort-Wort-Übersetzung anpassen zu können, weil es im bilingualen Wörterbuch keine oder nur wenige Nomenphrasen gibt. Sie haben berichtet, dass ihr Kombinationsmodell die Leistung besser war als hochwertige maschinelle Übersetzungen. [RAPP99] hat ein einfaches maschinenlesbares Wörterbuch mit der Kookkurrenzhäufigkeit der im Korpus vorgekommenen Wörter kombiniert. Dazu wurde ein Wörterbuch aus englisch-deutschen Korpora erzeugt, die aus vergleichbaren, aber nicht analogen Texten bestanden.

3.1.3 Vergleichsmaß

Weil die Ausgangsliste der Dokumente der krosslingualen Informationsrückgewinnung aus den Dokumenten der verschiedenen Sprachen in der Reihefolge der Bewertung besteht, können die relevanten Dokumente durch die Precision und den Recall sowie durch normale monolinguale Informationsrückgewinnung gemessen werden (siehe Abschnitt 2.3.1).

3.1.4 Ressource

Für die krosslinguale IR-Forschung braucht man einige wichtige Ressourcen, um das System zu konstruieren oder zu evaluieren. [GONZ00] hat die linguistische Ressource für CLIR in drei Gruppen unterteilt:

1. Wörterbuch bzw. Übertragungsmethode
2. Ausgerichtete Korpora
3. Natürliche Sprachverarbeitungsanwendung

Gonzalo berichtet in CLEF 2000, dass die teilnehmenden Gruppen ein Problem in der Güte der von CLEF gelieferten Ressourcen sehen. Dieses Problem ergibt sich daraus, dass die kostenlosen Ressourcen im Vergleich zu den nicht kostenfreien Ressourcen ei-

ne geringere Qualität aufweisen. Außerdem spiegelt sich im Ergebnis der Evaluierung auch die Qualität der Ressourcen wider. Man kann also sagen, dass die Leistung des Systems sowohl von der verwendeten Methode als auch von der Güte der Ressourcen abhängt.

Das Wörterbuch ist sehr wichtig, um die Anfragen zu übersetzen. Leider decken die meisten Wörterbücher aber nicht alle Fachgebiete ab, sondern liefern nur allgemeine semantische Bedeutungen. Das kostenlose Wörterbuch, das man im Internet herunterladen kann, enthält nicht sehr viele Wörter und besitzt zusätzlich die obengenannten Nachteile. Deswegen betreiben viele CLIR-Gruppen ihre Forschungen rein mit der Korpusbasis.

Um das Wörterbuch aus dem Korpus bzw. der Korpusbasis zu konstruieren, hängt alles von den parallelen oder vergleichbaren Korpora ab. Aus diesem Grund stellen die parallelen Korpora eine wichtige Ressource dar. Der Vorteil der Korpusbasis ist die Fähigkeit, „eigene“ Fachwörter zu übersetzen. Wenn es genug Korpora gibt, passt sich die Korpusbasis dem Fachgebiet an. Im Gegensatz dazu entsteht eine deutliche Beschränkung aufgrund des Mangels an der parallelen und vergleichbaren Korpora. Die Qualität des Korpus bzw. der Korpora ist ebenfalls sehr wichtig, weil Fehler aller Art, z.B. Buchstabierfehler, Satzbaufehler oder Semantikfehler usw. die Güte der datenorientierten Methode deutlich verringern können [SCI97].

Man verwendet die vergleichbaren Korpora nicht nur um das Korpusbasiswörterbuch zu konstruieren, sondern auch um das Muster der Phrase zu ermitteln (vgl. [BALL00] und [GNXZZH98]). Außerdem sind die vergleichbaren Korpora für die statistische sprachliche Verarbeitung sehr wichtig.

In der Realität kosten parallele oder vergleichbare Korpora sehr viel und man findet diese nur in geringer Zahl. Die Idee eine Wörterbuchbasis und eine Korpusbasis zu integrieren ist sicherlich von Vorteil [GONZ00].

Natürliche Sprachverarbeitung ist bisher eine bekannte Methode, die in vielen Forschungen verwendet wird. Die Stammform-, Part-of-Speech und Morphologieanalysen

werden zum Teil in die Entwicklung monolingualer und multilingualer IR Systemen eingebracht. Mit der natürlichen Sprachverarbeitung werden einige sprachliche Probleme vermieden, insbesondere in nicht-englischen Sprachen.

3.2 Frühe Ansätze der Bilingualen Suche

Die heutigen Forschungen bezüglich des krosslingualen Informationsretrievals können in zwei unterschiedliche Gruppen unterteilt werden. Die erste Gruppe bezieht sich auf die Forschungen, die sich mit der technischen Entwicklung und einer hochwertigen Ressourcenvorbereitung beschäftigen, damit eine gute Leistung des Systems gewährleistet werden kann. Die zweite Gruppe beinhaltet die Verständnisforschungen, die sich mit dem Bedürfnis des Nutzers und der realen Nutzung beschäftigen, damit das Verhalten des Nutzers verstanden und die Verbindung zwischen dem ergonomischen System und dem Nutzer vorbereitet werden kann.

Die wichtigste Aussage in der Arbeit von Douglas W. Oard [OARD97b] ist das reale Problem des normalen Suchers. Dieses Problem besteht darin, wie der Nutzer die benötigten Dokumente aus der Rangliste am besten auswählen soll, wenn er eine andere Sprache nicht so gut verstehen kann.

In CLEF 2000 wurde das künftige Ziel der CLIR-Entwicklung unter der Bedingung der grundsätzlichen Verbindung von leistungsfähigen monolingualen Informationsretrievalsystemen zusammengestellt. Wegen der ungenügenden Leistungsfähigkeit des IR-Systems in einigen Sprachen muss das IR-System verbessert werden. Aus diesem Grund kann man die monolingualen IR-Forschungen in CLEF wiederfinden. In einigen Sprachen findet man sehr wenige Ressourcen. Dadurch ist die Fortentwicklung erheblich behindert. Es wurden ein paar Beispiele in [GKP02] aufgeführt, z.B. dass es in Tamile und Zulu sehr wenige Ressourcen gibt, obwohl eine nicht geringe Anzahl von Menschen diese Sprachen benutzen. Das Bedürfnis an krosslingualer Information resultiert aus kulturellen und medizinischen Faktoren. Der Unterschied der herkömmlichen sprachlichen Wurzeln steht ein großes Problem dar, weil speziale Techniken auf unter-

schiedlichen Sprachfamilien möglicherweise nicht gleichermaßen greifen. Die unterschiedlichen Vorläufer der Sprachen verlangen eine eigene Technik und eine genaue Analyse.

Man kann den Nachteil der mangelhaften Ressourcen in einigen Sprachen an dieser Stelle deutlich sehen, da ohne effiziente Ressourcen eine große Behinderung der krosslingualen IR-Forschung entsteht. Bisher gibt es die Erfolge in dem krosslingualen Informationsretrieval fast nur mit der englischen Sprache, beispielsweise englisch-deutsch, englisch-französisch, englisch-italienisch, englisch-japanisch oder englisch-chinesisch, weil die Entwicklung noch von den sprachlichen Ressourcen[15] abhängt.

Um die wichtigen Probleme der krosslingualen Suche, nämlich die Mehrdeutigkeit und die mangelhaften Ressourcen, zu vermindern, werden von IR-Forschern verschiedene Übersetzungstechniken mit eigenem IR-System integriert. Man benutzt zusätzlich die natürliche Sprachverarbeitung, um den Sinn der Suchanfrage zu übertragen. Beim [GNXZZH98] werden die statistische Korpusbasis und Wörterbuchbasis angewendet, um die einzelnen Wörter und Mehrwortgruppen geeignet zu übersetzen, damit die Semantik der Mehrwortgruppe gewahrt werden kann. Die Notwendigkeit der semantischen Sinnbewahrung[16] erfordert die Entwicklung neuer leistungsfähiger Techniken für bilinguale bzw. multilinguale Suche. [FLUH04] hat sein multilinguales Informationsretrieval entworfen, indem die Sprachen durch die konzeptionelle Maschinenübersetzung verbunden werden, um die Leistung der Übersetzung der Mehrwortgruppen zu verbessern. Durch die Erweiterung der Suchwörter des Thesaurus und der eng verwandten Terme in [GEJI99] wird gezeigt, dass durch die speziellen Wortschatzeigenschaften des Thesaurus mehr als eine doppelte Precision des Retrievals gegenüber der universellen maschinellen Übersetzung erreicht wird und die eng verwandten Terme die Leistung des Thesaurus herabsetzen können. [LALI90] und [DLLL97] haben eine alternative

[15] Beispielsweise paralleles Korpus, maschinelle Übersetzung.

[16] Im Gegensatz zu der einfachen Übersetzung

Strategie präsentiert, um den Mangel an einem guten Wörterbuch zu umgehen. Bei ihr wird die Information in dem Konzeptraum durch den Latent-Semantik-Index transformiert. Beim MIETTA-Projekt [BNX98] werden die multilingualen Indexe in einem strukturellen Baum an der Stelle eingebaut, wo dieselbe konzeptionelle Klassifikation vorliegt. Das Eurospider-Projekt [BKSK99] hängt vom Mangel an Wörterbüchern in einigen Sprachpaarungen ab und bedient sich des Prinzips der Umsetzungssprache. Trotz der Arbeit von Rapp [RAPP99] handelt es sich nicht direkt um ein krosslinguales Informationsretrieval, obwohl seine Technik der Übersetzung unter Mithilfe des elektronischen Wörterbuchs und der Wortkookkurrenz für die krosslinguale Suche nützlich sein kann. Die Idee einer Übersetzung durch die Kookkurrenz wird auch [GNXZZH98] für englisch-chinesisch benutzt.

Andere Probleme tauchen bei asiatischen Sprachen auf, z.B. Wortgrenzenerkennung oder transliterierte Wörter. Dadurch ist die CLIR-Entwicklung in einer asiatischen Sprache von sprachlicher Vorverarbeitung abhängig. Daher bemühen sich zahlreiche Forschungen zunächst um gründliche Informationsvorbereitung. [SUPR99] verwendete phonetische Kodierung, nämlich Soundex-Kodierung von Odell und Russel, um das Problem von transliterierten Wörtern zu vermeiden. [JARU01] benutzt SWATH (Smart Word Analysis for THai), um die Sätze abzutrennen. Außerdem wird ein maschinell lesbares Wörterbuch verwendet, um die Anfrage zu übersetzen.

Bislang gibt es wenige Arbeit im Bereich der bilingualen Suche, die auf der Konzeptübertragung mit Hilfe der Mensch-Maschine-Schnittstelle basiert. Eine, die wir im Netz gefunden haben, ist die bilinguale InfoMap[17]. Die Grundidee in diesem Projekt ist, dass die ähnlichen Begriffe in unterschiedlichen Sprachen in demselben semantischen Konzeptraum abgelegt werden. Die Relation zwischen Begriffen wird durch die Assoziation zwischen den umgebungsbedingten und festgelegten Wörtern erzeugt, indem die parallelen Dokumente in einem Dokument verknüpft werden [WIDO02]. Durch die grafische

[17] http://infomap.stanford.edu/

Darstellung kann der Nutzer die entsprechenden Begriffe beider Sprachen im Zusammenhang sehen und die entsprechenden Dokumente finden. Die Visualisierung der semantischen Informationen basiert auf der Arbeit von D. Widdows et al [WCD02]. Die Gruppen der semantischen Informationen teilen sich in Cluster als Untergraph durch die Abtrennung des Kernwortes. Die hinterstehende Technik ist der Latent-Semantik-Index auf der Term-Term-Matrix gemäß Inhalt-Blocklagerung-Methode und die Graphtheorie.

Obwohl englisch-deutsch parallele Korpora in unser Projekt mit einbezogen werden, nehmen wir an, dass die präsentierte Methode auch für ein anderes sprachliches Paar funktionieren könnte. Es soll erst einmal untersucht und sichergestellt werden, dass unsere Methode unter einfachen Bedingungen zufriedenstellend funktioniert.

3.3 SENTRAX für CLIR

Die SENTRAX ist eine leistungsfähige (monolinguale) IR-Anwendung (vgl. Abschnitt 2.5). Durch die Funktionen der SENTRAX können unterschiedliche Arten Informationen mit dem Nutzer ausgetauscht werden. Sie erlaubt während des Suchprozesses beispielsweise eine Auswahl durch Interaktion mit dem Nutzer. Ein besonderes Merkmal der SENTRAX, das sie als ein gutes Werkzeug für ein krosslinguales IR-System qualifiziert, ist das Konzeptnetz, das durch Kookkurrenzhäufigkeiten erzeugt und durch Clusterverfahren gruppiert wird. Die Funktion „ContextMap“ der SENTRAX zeigt die verwandten Wortgruppen, die von einem Kontextfenster in der Sammlung gefunden werden. Aus diesem Grund und der Hypothese, dass parallele oder vergleichbare Dokumente in unterschiedlichen Sprachen eine ähnliche Struktur der semantischen konstruierten Wörter haben könnten, kann vermutet werden, dass die SENTRAX das Leistungsvermögen eines krosslingualen IR-Systems erheblich steigern könnte.

Die Wortbeziehung ist eine Hilfe, um die semantische Bedeutung des fokussierten Wortes zu bestimmen, weil ein Wort häufig nicht nur eine Bedeutung hat. Seine bestimmte Bedeutung hängt von seiner Umgebung ab, in der es verwendet wird. Das englische

Nomen „the bank“ hat z.B. in verschiedenen Umgebungen unterschiedliche Bedeutungen:

Umgebung	Mögliche Bedeutung und ggf. Übersetzung
Finanzen	„die Bank“ „das Bankinstitut“ „das Kreditinstitut“
Geographie	„die Schräglage“ „die Anhöhe“ „der Damm“
Sport	„die Bande des Billardtisches“

Analog dazu hat aber auch „die Bank“ in verschiedenen englischen Umgebungen andere Bedeutungen:

Umgebung	Mögliche Bedeutung und ggf. Übersetzung
Finanz	„the bank“
Technik	„the plate“
Sache	„the bench“

Wenn das Wort ohne seine Umgebung vorliegt, weiß man nicht exakt, welchen Sinn das Wort besitzt. Erst wenn man es im Zusammenhang mit den anderen Wörtern, mit seinem Kontext, betrachtet, kann man ihm eine deutlich exaktere, idealerweise eindeutige, Bedeutung zuweisen.

Der Kern des CLIR-Verfahrens ist üblicherweise die Übersetzung der Anfrage. Eine Ablenkung von den relevanten Dokumenten kann aufgrund der Mehrdeutigkeit der Übersetzung entstehen, insbesondere bei der Erweiterung der Anfrage. Da die SENTRAX die datenorientierten Begriffe in Bezug auf die Anfrage anbietet, um die Anfrage zu erweitern, ist sichergestellt, dass die Erweiterung nur auf die existierenden Domänen beschränkt bleibt und die Mehrdeutigkeit vermindert werden kann. Außerdem könnte die Wort-Wort-Übersetzung, die auf reiner Wortebene stattfindet, zwei von drei Fehlern der krosslingualen Übertragung verursachen (siehe Abschnitt 3.1.2). Die Bearbeitung auf der Konzeptebene sollte vielleicht solche Fehler vermeiden. Durch Neugruppierung[18] von Attributen verschiedener Konzepte entsteht ein neues Konzept, welches weniger Übersetzungen erlaubt, als durch Wort-Wort-Übersetzung denkbar wären. Bei der SENTRAX werden alle mögliche Übersetzungen der gewählten Begriffe direkt in die

[18] Durch benutzerdefinierte Auswahl von Begriffen aus der ContextMap

andere Sprache übertragen, gegenwärtig noch manuell. Durch den Vergleich der ContextMaps beider Sprachen kann der Benutzer auch bei Unkenntnis einiger Fremdwörter die Begriffswolke auswählen, die seiner Meinung nach der ursprünglichen am nächsten kommt. Auf diese Weise, so darf vermutet werden, führt das zu einem parallelen Treffer. Eine Strategie zur Automatisierung des Konzeptabgleichs für die SENTRAX wird in den Abschnitten 4.2.4 - 4.2.5 beschrieben.

Als Vorteile der SENTRAX in der Verwendung zur krosslingualen Suche bieten sich folgende an:

- Die mehrdeutige Übersetzung ist ein großes Problem. Die Wörter haben ihre eindeutige Bedeutung nur im Zusammenhang mit ihrem Kontext. Um die Mehrdeutigkeit zu vermeiden, wird das Konzept, das von einem Kernwort und dessen verwandten Wörtern gebildet wird, im Gegensatz zu der einfachen Wort-Wort-Übersetzung, in ein anderes sprachliches System übertragen. Eine solche konzeptionelle Übertragung durch die SENTRAX könnte die gesamte semantische Bedeutung erhalten, weil der SENTRAX Index aus der Wortbeziehung im Kontext erschaffen wird und den Nutzer angezeigte Wörter auswählen lässt. Durch die konzeptionelle Kombination der ausgewählten Wörter kann die semantische Bedeutung des Ausgangskonzeptes eingeengt werden.
- Obwohl die vom Nutzer getroffene Wortauswahl hilft, sein Konzept zu verwirklichen, sind die Synonyme aus der Übersetzung weiterhin ein Problem. Ein Wort einer Sprache kann viele Übersetzungen in der anderen Sprache repräsentieren. Diese Wörter sind eventuell Synonyme, die Homonymie oder die Polysemie. Der Zusammenhang zwischen dem übersetzten Wort und seiner Umgebung kann uns verraten, welche Übersetzung das ursprüngliche Wort am besten trifft. Mit der Übertragungsmethode und Vergleichstrategie könnte die ähnlichste vertretene Wortstruktur gefunden werden. Dadurch werden die geeigneten Wörter, die das gleiche Konzept in der ursprünglichen Sprache repräsentieren könnten, von dem System automatisch ausgewählt. Die Übertragungsmethode und die Vergleichstrategie finden sich in den Abschnitten 4.2.3-4.2.5.

- Wie es bei Douglas W. Oard [OADO96] erwähnt wird, ist es für jene nicht einfach, die ganze Retrievalliste durchzulesen, die die fremde Sprache nicht so gut beherrschen. Bei der SENTRAX muss man nur die wichtigen Wörter verstehen, um den Schlüssel zu wählen. Dieser Schlüssel gehört zu einer bestimmten Tür, hinter der sich die zum aufgebauten Schlüsselkonzept verwandten Dokumente befinden. Obwohl bei der üblichen Anfrageerweiterung die Priorität der zusätzlichen Wörter geringer ist als bei den anfänglichen Suchwörtern, werden die ausgewählten Zusatzterme bei der SENTRAX mit gleicher Priorität betrachtet, um das Konzept einzuengen.
- Das Konzeptnetz, das durch die ContextMap-Funktion erzeugt wird, ist ein von einer statistischen Methode erschaffener „Ähnlichkeitsthesaurus“ (siehe Abschnitt 2.4.1.3). Weil das Konzeptnetz der SENTRAX gegenwärtig allein auf dem Korpus basiert bzw. datenorientiert ist, können die angebotenen verwandten Begriffe auch in einem der Texte gefunden werden. Dies ist also anders als bei der Anfrageerweiterung mittels linguistischem Thesaurus, da dort nicht sichergestellt werden kann, dass die erweiterten Terme im Korpus irgendwo auftauchen.
- Ein Ansatz Ausdrücke abzugleichen ist das LSI-Modell, das in dem IR-System verwendet wird. Die Idee des LSI-Modells ist, dass man die Ausdrücke zwecks Nutzung des Konzepts abgleichen kann, indem man annimmt, dass die resultierenden Dimensionen die Basiskonzepte sprachlich unabhängig repräsentieren. In der SENTRAX werden die Ausdrücke ebenfalls als einzelnes Konzept von einer in die andere Sprache übertragen. Der Unterschied besteht darin, dass das LSI-Modell die Dimensionen von Ausdrücke-Dokumenten zum Konzept reduziert, während die SENTRAX die Attribute von einem Nutzer als Konzept verwendet. Der Nachteil beim LSI-Modell besteht darin, dass der optimale Konzeptraum aus dem Experiment gebildet werden muss.
- Bei [BNX98] wird der konzeptionelle klassifizierte Baum verwendet. Die Vorverarbeitung, um die multilingualen Indexe auf einem geeigneten Knoten abzulegen, muss unbedingt durchgeführt werden, was die Übersetzung aller Doku-

mente einschließt. Bei der SENTRAX brauchen die Dokumente nicht erst übersetzen zu werden. Sie erzeugt den statistischen Zusammenhang der Wörter in der zugehörigen Sprache. Ähnliche bilinguale Strukturen der Wortsbeziehungen werden bei unserer bilingualen Suche berechnet, um die Konzepte zu vergleichen. Somit werden die übersetzten Suchwörter und ihre Zusatzbegriffe in selben Sinne wie bei den Ausgangswörtern interpretiert.

- In [BALL00] und [GNXZZH98] wird die Phrase durch das Muster des Zusammentreffens ermittelt, wofür man die vergleichbaren Dokumente benötigt. In der SENTRAX ist dies anders, weil sie auf statistischem Zusammentreffen der Ausdrücke basiert. Der Vorteil ist, dass die SENTRAX keine vergleichbaren Dokumente benötigt, da diese meist wenig vorhanden sind.
- Obwohl die InfoMap eine ähnliche grafische Darstellung wie SENTRAX produziert, entsteht die Relation nur aus der direkten Assoziation zwischen den festgelegten Nomen und allen Nomen in Korpus. Eine solche Relation hat aber keinen semantischen Zusammenhang. Das semantische Netz wird durch die Assoziation der Nomen in bestimmten Mustern mit Hilfe der Graphtheorie getrennt aufgebaut. Die Wortrelation der SENTRAX ist aber von der direkten und indirekten Assoziation der Wörter zusammengestellt worden, ohne das eindeutige Muster zu erkennen. Ihre grafische Darstellung stammt aus der Cluster-Methode und Singularwertzerlegung.
- Die Methode von InfoMap ist für die bilinguale Anwendung vom abgestimmten parallelen Korpus abhängig. Bei der SENTRAX gilt dies nicht. Ihr semantisches Konzeptnetz kann auch verglichen werden, wenn es auch dem nicht parallelen Korpus erzeugt wird.

4 UNSER ANSATZ DER BILINGUALEN SUCHE

4.1 Grundidee

Da der wichtige Anschluss der krosslingualen Suche die Anfrageübertragung ist, beschäftigen sich viele Forscher, um den gleichmäßigen Sinn von der Anfrage aufzubewahren. Die Übertragung der konzeptionellen Anfrage mittels Konzeptnetz für krosslinguale Suche durch die Funktion von SENTRAX basiert auf dieser Idee. Die Semantik des Konzepts wird durch die gemeinsame Übersetzung der von Nutzer ausgewählten Eigenschaften festgehalten. Somit sind die Anfrage in der Fremdsprache derselbe bzw. ähnlichste Sinn wie im ursprünglichen Suchsystem.

4.1.1 Semantische Stammformreduzierung

Vor zwanzig Jahren bemühten viele IR-Forscher mit der konventionellen statistischen Methode, um die Indexierung zu verbessern. Viele untersuchten die Anfrage mit Hilfe eines Wörterbuchs bzw. Thesaurus, um diese zu erweitern, damit sie bessere Precision- und Recallwerte erhalten. Vor ca. einem Jahrzehnt vertiefte sich die neue Generation der IR-Forschung in der Sprachtechnik. Die Indexierung und die Anfrage sind wichtige Faktoren, die in der IR-Richtung beachtet werden müssen. Dabei stellen sich zwei Fragen:

1. Welches Wort soll indexiert werden?
2. Welche Suchwörter sind gut, um die Absicht des Nutzers zu repräsentieren?

Diese beiden Fragen stehen in engem Zusammenhang, sie sind eng miteinander verwandt. Die indexierten Wörter könnten dem angefragten Wort angepasst werden, das bei der Suchanfrage eingegeben wird. D.h. es wird angenommen, dass das indexierte Wort den Suchbegriff bzw. der Anfrage ähneln soll.

Um die zweite Frage zu beantworten, bewebt sich die SENTRAX dafür. Es wird angenommen, dass eine Suchanfrage, die durch einen Nutzer hingegeben wird, der nicht an das Suchprogramm gewöhnt ist, zu den ungewünschten Dateien einführt. Durch die ContextMap würde die Anfrage erweitert, damit der Sucher eine entsprechende Wortgruppe bekommt und seine Idee mit den erweiternden Wörtern besser ausdrücken kann. Dank der Morphologie kann die Worttype sowohl in einer englischen als auch in der deutschen Datei erkannt werden. Mit Hilfe der *POS-Markierungen* können die Nomen bzw. andere bedeutungstragende Worttype einfach aus dem Dokument herausgenommen werden. Die Lemma-Funktion von einiger morphologischen Anwendung ist sehr behilflich, um die Flexion und die Derivation zur ihren Stammform reduzieren zu können.

Inzwischen dem Prozess der ContextMap müssen alle Wörter durch die direkte und indirekte Assoziationen geführt werden [ACKE00]. Das Verb „geben“ besitzt beispielsweise verschiedene Konjugationsformen, z.B. „gibt“, „gebe“, „gab“, „gegeben“. Ebenso hat das Nomen „Buch“ unterschiedliche Deklinationsformen, z.B. „Bücher“, „Büchern“. Die SENTEXT unterscheidet die Wörter „Buch“ und „Bücher“ sowie „geben“, „gegeben“ und „gab“. Das heißt, es können verschiedene Sätze wie die folgenden kreiert werden:

„Michael gibt mir das Buch.“ und „Michael gab mir das Buch.“

„Michael gibt mir ein Buch.“ und „Michael gibt mir zwei Bücher.“

Es könnte also passieren, dass die ähnlichen Sätze wie beispielweise diese vier Sätze in einer Datei vorkommen. Werden nun vier *Großfenster* für die unterschiedlichen Formen verwendet, so ließen sich im Zusammenhang der *assoziierten ersten Ordnung bzw. der direkten Assoziation* die Verbindungen zwischen „Michael“, „gibt/gab“ und „Buch/Bücher“ herleiten. Diese sind bei dem ersten Satzpaar „Michael—gibt“, „Michael—gab“, „gibt—Buch“ und „gab—Buch“, und bei dem zweiten Satzpaar sind es „Michael—gibt“, „gibt—Buch“ und „gibt—Bücher“.

Durch die *assoziierte zweite Ordnung bzw. die indirekte Assoziation* wird der eigentliche Zusammenhang zwischen „Michael“ und „Buch“ für das erste Satzpaar durch

„gibt“ bzw. „gab“ hergestellt. Es gibt kein Problem in dieser Situation, weil die Assoziationen zwischen „Michael“ und „Buch“ durch alle assoziierten Wörter in der 2.Ordnung summiert werden können. Es spielt kaum eine Rolle, ob sie durch „gibt“ oder „gab“ verbunden werden. Betrachtet man die Gleichung der Assoziation, so erkennt man, dass die Ähnlichkeit zwischen „Michael“ und „Buch“ durch

$$sim(M,B) = \frac{ass(Michael, gibt)\cdot ass(gibt, Buch) + ass(Michael, gab)\cdot ass(gab, Buch)}{\left[ass(Michael, gibt)^2 + ass(Michael, gab)^2\right]^{1/2}\cdot\left[ass(gibt, Buch)^2 + ass(gab, Buch)^2\right]^{1/2}}$$

berechnet werden kann. Nehmen wir an, dass $ass(M, g_i)$ und $ass(g_i, B)$ jeweils „0,1“ betragen. Daraus folgt, dass die Ähnlichkeit zwischen „Michael“ und „Buch“ 0,0224 beträgt. Falls die morphologische Methode genutzt wird, um die Stammform des Verbs zu erkennen, wären „gibt“ und „gab“ nicht zu unterscheiden. Mit der Stammform werden die Assoziation $ass(M, geben)$ und $ass(geben, B)$ „0,2“ sein, weil sie gleich zweimal zutrifft. Durch diese Änderung ergibt sich als Ähnlichkeit zwischen „Michael“ und „Buch“ der Wert „1“. Dies wird deutlich, da gilt: $\sum (a_n)^2 > \left(\sum a_n\right)^2$. Dadurch entsteht ein wesentlicher Unterschied. Aus der Umformung des Verbs zur Stammform erhält man eine wesentlich stärkere Ähnlichkeit. Dieser Einfluss hat eine große Auswirkung auf die Abbildung der Beziehung der Wörter im Erzeugungsprozess der ContextMap. Die assoziierten Nomen, die durch das gleiche konjugierte Verb verbunden wurden, werden verstärkt und steigen deutlich in der Rangliste. In besonders für deutsche Sprache entsteht so oft diese Situation wegen der Konjugation. Auf die gleiche Art kann mit Hilfe der Morphologie die Stammform von „Bücher“ ermittelt werden. Es ist klar, dass die Beziehung zwischen „Michael“ und „Buch“ verdoppelt wird, wenn man „Buch“ statt „Bücher“ verwendet. Wenn viele solche Paare assoziierter Nomen in einem Text vorkommen, kann man die Beziehungsgrafik dieses Ereignisses (siehe Abbildung 7) abbilden. Darüber hinaus wird das Nomen auf die Stammform reduziert, um weniger Platz in der SENTRAX zu benötigen, denn der Benutzer interessiert sich nicht dafür, ob das Nomen im Singular oder im Plural vorkommt. Es ist offensichtlich, dass auf diese Weise Speicherplatz und Rechenzeit reduziert werden können. Der Vorteil von der Nutzung der Morphologie liegt somit klar auf der Hand.

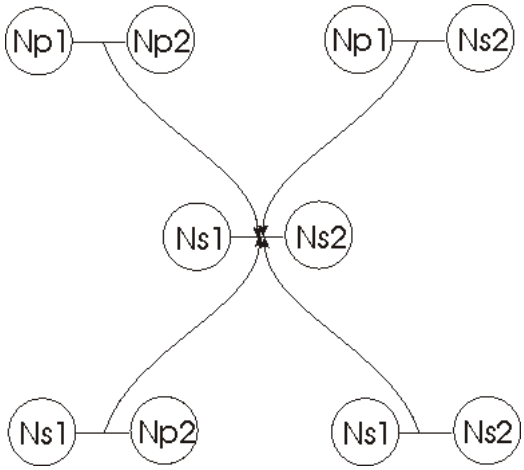


Abbildung 7 Die Grafik zeigt die Verringerung der Beziehungen zwischen dem Nomenpaar. Die Verbindungen zwischen den Kreisen repräsentieren die Gewichtstärke. „Ns“ repräsentiert Nomen im Singular, „Np“ Nomen im Plural.

4.1.2 Suche durch Konzept

Zu einer einfachen Frage, „was ist typisches deutsche Essen?“, denkt man daran, dass die allgemeine Antwort das Essen ist, das aus Deutschland stammt. Natürlich ist die Antwort von der Region abhängig. Vielleicht versteckt auch der Sinn von „weltbekannt“, „die Art“ und „das kulinarische Zeichen“. Man versucht die erwartete Antwort durch eine Gruppe von Eigenschaften zu beschreiben. Diese Gruppe von den Eigenschaften kann ein Konzept von einem Sucher bilden. Um das Konzept zu erfüllen, können entweder die Beispiele oder zusätzliche Eigenschaften auch dazu eingebracht werden.

Das obige Beispiel zeigt sich, wenn man die Antwort von einer Frage benötigt, versucht man zuerst eigenes Konzept der suchenden Antwort zu bilden. Dann werden die Eigenschaften bzw. Attribute des Konzepts beschieben. Die Gruppe der Eigenschaften kann seitens des Suchers sein Konzept vertreten. Sie sind vom Sucher abhängig und können unterschiedlich sein. Aber sie sollen zu der richtigen Antwort einführen. Die solche Eigenschaften könnten in der Wahrheit noch andere Konzepte besitzen. Die konzeptionelle Ablenkung könnte passieren, weil die vertretenden Eigenschaften unklar sind. Die

Lösung dieses Problems ist, dass das Konzept durch die zusätzlichen datenorientierten Eigenschaften verschärft wird.

Beim unseren krosslingualen Suche-Ansatz wird das Konzept nicht nur durch die Eigenschaften vom Anfangsystem zum Ausgangsystem übertragen, sondern auch ihre Beziehungen. Mithilfe der Kookkurrenz können die übersetzten Eigenschaften dasselbe Konzept bewahren (siehe Konzeptnetz im Abschnitt 2.5.1.3), indem die Homonymie, Polysemie und Synonyme der Übersetzungen miteinander kontrolliert werden. Das Gleichgewicht der gesamten Kookkurrenz kann nicht nur die besten Übersetzungen bringen, sondern auch das Konzept behalten.

4.2 Technische Voraussetzungen

4.2.1 Vorverarbeitung

Da die SENTRAX auf der Wortkookkurrenz basiert und jede Sprache eigenen Charakter hat, ist es sehr wichtig überflüssige Ausdrücke herauszufiltern, ohne dabei wichtige Information zu verleeren. Das Ziel der Reduzierung ist der Ausgleich der Sprache. Wenn der Text nur informationstragende Wörter hätte, würde die Information durch die Kookkurrenz der uninformatorischen Ausdrücke nicht streuen. Nun bringen aber z.B. Flexion eine Abweichung in der Wortbeziehung mit sich (vgl. Abschnitt. 4.1.1), was besonders in der deutschen Sprache vorkommt. Auch das „Kompositum“ taucht häufig im deutschen Text auf, beim Englischen hingegen die „Mehrwortgruppe“. Die deutsche Genitivform wird sowohl ohne untergeordnete Konjunktion, beispielweise „die Verbesserung der tatsächlichen Qualität der gebotenen Bildung“, als auch mit untergeordneter Konjunktion, beispielweise „die Verbesserung von tatsächlicher Qualität der gebotenen Bildung“, geschrieben, allerdings wird in englisch „improvents in the quality of educa-

tion provided“[19] geschrieben. Diese Beispiele zeigen den Nutzungsunterschied zwischen der deutschen und englischen Sprache, der durch die Kookkurrenz die Wortbeziehungen direkt bewirkt. Abgesehen davon kommt auch eine große Menge an Beziehungen zwischen Wortpaaren vor, wenn es viele uninformatorische Wörter im Korpus gibt. Ein erhöhter Zeit- und Speicheraufwand sind natürlich die Folge der unnötigen Beziehungen. Weil die SENTRAX Berechnungen teilweise in Echtzeit durchführt, ist die Verringerung der unbrauchbaren Wortarten nicht nur für sprachlichen Ausgleich wichtig, sondern auch für die Rechenzeit und den Speicheraufwand. Dieser Schritt kann mit Hilfe der Anwendung TIHO automatisch erfolgen [ZAND06], die parallel zu dieser Arbeit entwickelt wird (siehe Abschnitt 7.3.1)

Dank der Tagger-Anwendung kann die Wortart erkannt werden. Um die unbrauchbare Information aufzuräumen, wird die benötigen Wortarten zunächst definiert, welche Wortarten erhalten bleiben sollen (TIHO-Anwendung unter der Option „SavePattern“ seihe 7.3). Mit der Lemma-Funktion können die abgeleiteten Wörter zu den Stammformen reduziert werden. Die einfachen Algorithmen werden entworfen, um das trennbare Verb, die Mehrwortgruppe, das Kompositum und das englische Verb mit seinen weiteren Elementen zu erkennen.

[19] In parallelen Dateien durch die Übersetzung von EuroParl-Projekt 2.0.

4.2.1.1 Tagger-Anwendung: TreeTagger

Der TreeTagger ist eine bekannte morphologische Anwendung, die von Helmut Schmidt an der Universität Stuttgart entwickelt wurde. Er kann mit dem englischen sowie deutschen Text arbeiten. Mit „Decision Tree“ und der Markov-Methode kann durch den TreeTagger eine hohe Genauigkeit erlangt werden [SCHM94]. Als *Menge der Markierungsarten*, POS Tagset, verwendet Schmidt die von Penn-Treebank-Tagset und IMS-Stuttgart-Tagset. Der TreeTagger kombiniert die Anwendungs-, Parameter-, Tokens- Übersetzungs-, Abkürzungs- und Batchdatei.

Der TreeTagger bietet viele Funktionen (siehe Abschnitt 7.1). Die wichtigsten Funktionen, die im unseren Ansatz genutzt werden können, sind der POS-Tagger und die Stammformanerkennung. Die Stammformanerkennung erfolgt durch die Option „-lemma“. Weil der TreeTagger nur ein Wort pro Zeile bearbeiten, lässt die Option „-token“ mit der perl-Anwendung bearbeiten.

Die untere ausgeschnittene Tabelle ist der Aussicht der Taggerdatei vom Text:

„...Das internationale europäische Jugendtreffen der ökumenischen Taiz-Gemeinschaft ist am Neujahrstag nach viertägiger Dauer mit Gebeten, Meditationen und einem Friedensappell zu Ende gegangen. 80000 junge Menschen aus 17 Ländern hatten an dem Treffen teilgenommen. Der Gründer der Taiz-Gemeinschaft, Bruder Roger Schutz, rief in einer Predigt alle Menschen zur Versöhnung auf. ...“

Das	ART	D
internationale	ADJA	international
europäische	ADJA	europäisch
Jugendtreffen	NN	Jugendtreffen
der	ART	D
ökumenischen	ADJA	ökumenisch
Taiz-Gemeinschaft	NN	<unknown>
ist	VAFIN	Sein
am	APPRART	Am
Neujahrstag	NN	Neujahrstag
nach	APPR	nach
viertägiger	ADJA	viertägig
Dauer	NN	Dauer
mit	APPR	Mit
Gebeten	NN	Gebet

,	$,	,
Meditationen	NN	Meditation
und	KON	und
einem	ART	Ein
Friedensappell	NN	Friedensappell
zu	APPR	Zu
Ende	NN	Ende
gegangen	VVPP	gehen
.	$.	.
80000	CARD	80000
junge	ADJA	jung
Menschen	NN	Mensch
aus	APPR	aus
17	CARD	17
Ländern	NN	Land
hatten	VAFIN	haben
an	APPR	An
dem	ART	d
Treffen	NN	Treffen
teilgenommen	VVPP	teilnehmen
.	$.	.
Der	ART	d
Gründer	NN	Gründer
der	ART	d
Taiz-Gemeinschaft	NN	<unknown>
,	$,	,
Bruder	NN	Bruder
Roger	NE	Roger
Schutz	NN	Schutz
,	$,	,
rief	VVFIN	rufen
in	APPR	in
einer	ART	ein
Predigt	NN	Predigt
alle	PIDAT	alle
Menschen	NN	Mensch
zur	APPRART	zur
Versöhnung	NN	Versöhnung
auf	PTKVZ	auf
.	$.	.

Tabelle 1 Die Wortartmarkierungsdatei: In der ersten Spalte stehen die originalen Wörter, in der zweiten die Wortarten und in der dritten die Stammformen.

4.2.1.2 Benötigen Wortartmuster

Benötigen deutsche Wortarten:

attributives Adjektiv (ADJA)
adverbiales oder prädikatives Adjektiv (ADJD)
Kardinalzahl (CARD)
Fremdsprachliches Material (FM)
normales Nomen (NN)
Eigennamen (NE)
substituierendes Indefinitpronomen (PIS)
attribuierendes Indefinitpronomen (PIDAT)
Relativpronomen substituierend (PRELS)
Relativpronomen attribuierend (PRELAT)
"zu" vor Infinitiv (PTKZU)
abgetrennter Verbzusatz (PTKVZ)
Kompositions-Erstglied (TRUNC)
finites Verb, voll (VVFIN)
Imperativ, voll (VVIMP)
Infinitiv, voll (VVINF)
Infinitiv mit "zu", voll (VVIZU)
Partizip Perfekt, voll (VVPP)
finites Verb, aux (VAFIN)
Imperativ, aux (VAIMP)
Infinitiv, aux (VAINF)
Partizip Perfekt, aux (VAPP)
Nichtwort, Sonderzeichen enthaltend (XY)
Komma ($,)
Satzbeendende Interpunktion ($.)
sonstige Satzzeichen; satzintern ($()

Benötigen englische Wortarten:

Adjective (JJ)
Adjective, comparative (JJR)
Adjective, superlative (JJS)
Noun, singular or mass (NN)
Noun, plural (NNS)
Proper noun, singular (NP)
Proper noun, plural (NPS)
Particle (RP)
Symbol (SYM)
Verb, base form (VB)
Verb, past tense (VBD)
Verb, gerund or present participle (VBG)
Verb, past participle (VBN)
Verb, non-3rd person singular present (VBP)
Verb, 3rd person singular present (VBZ)
Punctuation Tags # $ " ( ) , . : ``

4.2.1.3 Stammform reduzieren

Wie in der Tabelle 1 gezeigt ist die dritte Spalte die Stammformen. Nach der benötigen Wortartmusteranerkennung wird die Stammform in der Stammform-Datei geschrieben. Das originale Wort und die Wortart werden auch in der Wortreduzierungs-Datei bzw. Wortart-Datei auch angefügt. Eine Index-Datei wird nun mit Hilfe der SENTRAX erzeugt, um auf die Adresse des Stammworts zurückgreifen zu können. Die Wortreduzierungs-Datei und die Wortart-Datei sowie die Stammform-Datei werden in den nachfolgenden Prozessen (siehe Abschnitt 4.2.1.4 bis 4.2.1.8) mitgeteilt.

4.2.1.4 Kompositum erkennen

In diesem Algorithmus geht es darum, den Teil der deutschen Schreibenweise der Komposita voll zu verbinden. Sooft schreibt man das Kompositum als Aufzählung in

sparsamer Form, z.B. „Sport- und Rechenzentrum“. Mit diesem Prozess erlangt man die normale Form als Lösung, nämlich „Sportzentrum und Rechenzentrum“, weil nur „Sport-“ keinen Sinn im Wörterbuch gibt bzw. „Sport“ und „Sportzentrum“ unterschiedliche Bedeutung haben.

Originalwort	Tagger	Stammform
...	...	...
Anwaltsbüro	NN	Anwaltsbüro
sind	VAFIN	sein
am	APPRART	am
Neujahrstag	NN	Neujahrstag
Bundes-	TRUNC	Bundes-
und	KON	und
Reichsbahn	NN	Reichsbahn
privatisiert	VVPP	privatisieren
worden	VAPP	werden
.	$.	.

Tabelle 2 Beispiel des Kompositums „Bundes-“ und „Reichsbahn“

Voraussetzung

Wenn das Wort mit Bindestrich am Ende bzw. die Wortartmarkierung „TRUNC“ gefunden wird,

Gegeben seien i = 1, j=1

1. *Das ursprüngliche Wort ohne Bindestrich in der ersten Spalte wird in die Box[j] eingefügt. Die dazugehörige Adresse wird gebildet. Der Zeiger geht ein Wort nach rechts bzw. in die nächste Zeile der Tabelle.*
2. *Falls die Wortartmarkierung „TRUNC“ in der zweiten Spalte gefunden wird, wird j=j+1 gesetzt und zu (1) zurückgekehrt.*
3. *Falls die Nomenmarke(NN,NE) noch nicht gefunden wurde, geht der Zeiger ein Wort weiter, sonst stoppt der Zeiger und geht zu (2) zurück.*
4. *Sei S ein Teilstring des Stammwortes von der Stelle n-i bis n, wobei n die Anzahl der Buchstaben im Wort ist. Das Stammwort befindet sich in der dritten Spalte.*

5. *Es wird geprüft, ob S im Wörterbuch gefunden wird.*
6. *Wenn S nicht im Wörterbuch gefunden wird, wird i = i+1 gesetzt und zu (4) zurückgekehrt.*
7. *Wenn S im Wörterbuch gefunden wird, verbindet sich das Wort in der Box[j] mit S für alle j. Die Markierung wird durch „Nomen" ersetzt, sonst wird das neue Fenster geöffnet, um manuell zu korrigieren.*

4.2.1.5 Deutsche Mehrwortgruppen verbinden

In diesem Algorithmus geht es um das Problem, Mehrwortgruppe zu verbinden. Insbesondere gibt es im Text die Abfolge des Eigenamens z.B. „Joschka Fischer". Statt „Joschka" und „Fischer" in zwei Wörter zu trennen, verbinden sich die beide in einer Mehrwortgruppe, weil nur das Wort „Fischer" allein Unklarheit verursachen kann, ob das der Nachname oder der Beruftäter bedeutet. Beim deutschen Text wird die Mehrwortgruppe nicht oft vorgekommen. Meisten werden entweder als Kompositum oder mit dem Bindestrich geschrieben, beispielweise „Bundesverfassungsgericht" oder „SpaCAM-Technologie". Bei der Behandlung des Kompositums soll es in seine Bestandteile zerlegt werden. Weil das Ziel ist, dass die Wortstrukturen der beiden Sprache möglich gleich sind, werden die englischen Mehrwortgruppen verbunden, statt das Kompositum aufzubrechen. Dieser Umweg kann den Aufwand von einer besonderen natürlichen Sprachverarbeitung vermeiden. Deshalb kann die deutsche Mehrwortgruppe in der Form des regulären Ausdrucks als (Nomen|Eigenname)* einfach geschrieben werden.

Originalwort	Tagger	Stammform
...	...	...
Ansprache	NN	Ansprache
an	APPR	an
Jean-Jacques	NE	Jean-Jacques
Dessalines	NN	<unknown>

Tabelle 3 Beispiel der deutschen Mehrwortgruppe „Jean-Jacques Dessalines".

Voraussetzung

Wenn der Tagger bzw. die Markierung „NN“ oder „NE“ antrifft, läuft folgender Algorithmus ab:

1. *Falls die Markierung „NE“ oder das Stammwort in der dritten Spalte <unknown> ist, wird das originale Wort in der ersten Spalte durch die Stammform überschrieben.*
2. *Die Adresse des ersten Wortes wird beibehalten. Der Zeiger geht ein Wort nach rechts bzw. in nächste Zeile der Tabelle.*
3. *Falls das Wort nicht das Nomen(NN,NE) ist, wird die letzte Adresse zurückgeliefert und er geht nach (4), sonst wird die Stammform „<unknown>“ durch ihr Originalwort ersetzt, dann geht der Zeiger ein Wort nach rechts bzw. in die nächste Zeile der Tabelle und (3) wird wiederholt.*
4. *Die Wörter von der Anfangsadresse bis vor die Endeadresse werden verkettet. Die Stammform wird neu bestimmtt. Die Markierung setzt sich auf das Nomen „N“.*

4.2.1.6 Englische Mehrwortgruppen verbinden

Es gibt die Anfolge von Nomen im englischen Text häufiger als im deutschen Text. Die englische Mehrwortgruppe bzw. die Folge von Nomen kann in der Form des regulären Ausdrucks als

(Adjektiv | Nomen)*(Nomenpräposition)?(Adjektiv | Nomen)*Nomen

geschrieben werden[20]. Daraus folgt, dass das Nomen am Ende der Reihefolge zuerst erkannt werden muss. Danach werden die anderen vorderen Wortarten ermittelt, um die komplette Mehrwortgruppe zu verknüpfen.

Das Problem des sprachlichen Nutzungsunterschiedes sollte verringert werden z.B. „west europe“ und „Westeuropa“, indem die Mehrwortgruppe erkannt wird, damit die semantische Bedeutung von „west europe“ und „Westeuropa“ nachfolge abgeglichen werden kann. Es gibt noch viele andere Konstellationen von Adjektiven und Nomen, (Adj+Nomen), in der das Adjektiv nur eine Eigenschaft anzeigt z.B. „vernünftiger Mensch“ und „reasonable humans“.[21] Diese würde ohne eine richtige Analyse mit natürlicher Sprachverarbeitung ein Problem erzeugen. In dieser Arbeit wird dieses Problem vernachlässigt, weil eine Entscheidung darüber abhängig wäre von ebendieser natürlichen Sprachverarbeitung, die hier nicht integriert wurde.

Dadurch wird der reguläre Ausdruck der Mehrwortgruppe folgendermaßen umgeformt.

(Nomen)*(Nomenpräposition)?(Nomen)*Nomen

Das Nomen im obengenannten regulären Ausdruck schließt den Singular, den Plural und den Eigenname ein, wobei die Wortartmarkierungen (bzw. tags) für die Nomen als „NN“, „NNS“, „NP“ und „NPS“ sind. Die englische Nomenpräposition befindet sich im Anhang dieser Dissertation (siehe Abschnitt 7.4.1).

Wenn es, wie eigentlich erwünscht, möglich wäre, die richtige Reihefolge der Mehrwortgruppe mit der Nomenpräposition mit dem obengenannten regulären Ausdruck zu ermitteln, ist das Verhalten der SENTRAX diesbezüglich als problematisch anzusehen. Der Grund dafür ist, dass die Assoziationen der Wörter im Englischen durch die

[20] siehe http://www1.cs.columbia.edu/~min/research/termer/termerCIE.html für die Form der Mehrwortgruppe des regulären Ausdrucks und http://www.amk.ca/python/howto/regex/ beispielweise für die Definition des regulären Ausdrucks.

[21] Übersetzung durch http://world.altavista.com/ von deutsch ins englisch

SENTRAX in diesem Fall verschoben werden könnten. Deshalb wird die englische Nomenpräposition einfach vernachlässigt. Dann ließe sich der insoweit reduzierte Ausdruck mit der SENTRAX ohne Assoziationsverschiebungen verarbeiten.

Ein Fremdwort kann aber auch in englischem Text auftauchen. Als Beispiel der Eigenname einer Schule:

„Corona/NP del/FW Mar/NP High/NP School/NP“

Man erkennt an diesem Beispiel, dass es sinnvoll ist, das Fremdwort zwischen dem Eigennamen in einer Mehrwortgruppe einzubauen. Außerdem ist eine lange Mehrwortgruppe inklusive Nomen am Ende in manchem Fall nicht geeignet, weil durch die Verknüpfung der semantische Sinn verloren gehen kann, wie man z.B. an

„Newport/NP Beach/NP house/NN tonight/NN“[22]

Der reguläre Ausdruck der Mehrwortgruppe für die SENTRAX wird deshalb bezüglich der Wortartmarkierungen umgeschrieben.

$$(NP|NPS)^{+}(NN|NNS|FW)*(NP|NPS)$$

Originalwort	Tagger	Stammform
...	...	...
at	IN	at
Corona	NP	Corona
del	FW	del
Mar	NP	Mar
High	NP	High
School	NP	School
for	IN	for
a	DT	a

Tabelle 4 Beispiel der englischen Mehrwortgruppe „Corona del Mar High School“.

[22] Ausgeschnitten von dem durch den TreeTagger markierten Text

Voraussetzung

Wenn die Markierung „NP“ oder „NPS“ angetroffen wird und die nächste Markierung nicht „NP“ oder „NPS“ ist, läuft folgender Algorithmus ab:

1. *Falls das Stammwort in der dritten Spalte nicht <unknown> ist, wird das Originalwort in der ersten Spalte mit seiner Stammform überschrieben. Der Zeiger geht ein Wort nach links bzw. in die vorhergehende Zeile der Tabelle.*
2. *Falls das Wort ein Nomen (NP,NPS,NN,NNS) oder Fremdwort (FW) ist, wird dieses durch die Stammform ersetzt und das neue Nomen mit dem vorher gefundenen Nomen verkettet. Die nun überflüssige Zeile wird gelöscht. Sonst wird das Programm beendet.*
3. *Der Zeiger geht ein Wort nach links bzw. in die vorhergehende Zeile der Tabelle. Schritt (2) wird nun wiederholt.*

4.2.1.7 Deutsches trennbares Verb zum Infinitiv umformen

In diesem Algorithmus geht es darum, trennbare deutsche Verben zu verbinden. Im Hauptsatz schreibt man das trennbare Verb in getrennter Form. Wenn der getrennte Teil gefunden wird, sucht dieser Prozess die entsprechenden Teile und verbindet das Grundverb und den abgetrennten Teil. Eine Möglichkeit zur Verwechslung trennbarer Verben in normalen Sätzen besteht darin, dass das Verb in einem Nebensatz aber auch zwischen Klammern stehen kann. Wenn der getrennte Teil außerhalb der Klammerung steht bzw. nicht zum Nebensatz gehört, kann der Prozess den Satz zwischen den Klammern bzw. den Nebensatz überspringen. Im anderen Falle darf der Prozess nur auf dem Satz zwischen den Klammern bzw. in dem Nebensatz laufen. Mancher getrennte Teil mit der Markierung „PTKVZ“ ist vielleicht nicht ein Teil von einem trennbaren Verb, sondern nur eine Zirkumposition rechts (vgl. Abschnitt 7.1.4.1).

Originalwort	Tagger	Stammform
...	...	...
Wir	PPER	wir
fordern	VVFIN	fordern
diejenigen	PDAT	diejenigen
Mitgliedstaaten	NN	Mitgliedstaat
,	$,	,
die	PRELS	d
noch	ADV	noch
keine	PIAT	kein
ausreichende	ADJA	ausreichend
Fördergebietskarte	NN	<unknown>
eingereicht	VVPP	einreichen
haben	VAINF	haben
,	$,	,
auf	PTKVZ	auf

Tabelle 5 Beispiel des deutschen trennbaren Verbs „auffordern“

Voraussetzung

Wenn die Markierung „PTKVZ“ gefunden wird, läuft folgender Algorithmus ab:

1. *Die nächste Marke wird geprüft, ob sie „$,“, „$.“, „$(“ oder „KON“ ist. Falls nein, wird ganze Zeile gelöscht und der Prozess beendet.*
2. *Das Stammwort in der dritten Spalte wird in einer Variable namens „getrenntTeil“ abgelegt. Die dazugehörige Adresse wird registiert. Der Zeiger geht ein Wort nach links bzw. in die vorhergehender Zeile der Tabelle.*
3. *Falls die Marke „VVFIN“ gefunden wird, wird der getrennte Teil zusammen mit dem Stammverb insofern überprüft, ob getrenntTeil+Stammverb im Wörterbuch gefunden werden kann. Wenn ja, wird das Wort mit dem getrennten Teil kombiniert.*
4. *Falls die Marke „$(“ gefunden wird, kann diese auf zwei verschiedene Arten vorkommen. (a) Die erste Möglichkeit ist, dass eine Klammer auf angetroffen wird und der getrennte Teil vor der zugehörigen Klammer zu steht bzw. in der Form des regulären Ausdruckes ($()..(PTKVZ)($(). In diesem Fall kann der*

Prozess terminieren. (b) Die zweite Möglichkeit ist, dass eine Klammer zu gefunden wird. In diesem Fall läuft die Ziegeposition bis zu der vorhergehenden zugehörigen Klammer auf und noch eins nach vorne.

5. *Falls die Marke „$," gefunden wird und die vorhergehende Marke „VVFIN" ist , kann ein trennbares Verb auf zwei Arten vorkommen: (a) Als eine Folge von Verben, die in Form des regulären Ausdruckes ((VVFIN)($,))*(VVFIN)(KON)(VVFIN) geschrieben werden können. In diesem Falle werden alle Verben zusammen mit dem getrennten Teil darauf geprüft, ob sie im Wörterbuch stehen. Dasjenige Verb, das im Wörterbucht steht, wird mit dem getrennten Teil kombiniert. (b) Als Teil eines Satzes, in dem ein weiterer untergeordneter Nebensatz vorkommt, welcher mit dem regulären Ausdruck (KOUS|PRELS)..(VVFIN) beschrieben werden kann. In diesem Falle wird der Zeiger auf das zu betrachtende Wort auf das nächste vorhergehende Komma oder den Satzanfang gesetzt.*
6. *Falls die Marke „$." angetroffen wird, wird der Suchprozess beendet. Falls das trennbare Verb nicht gefunden wurde, wird die Markierung „PTKVZ" zu „APZR" geändert.*
7. *Sonst geht der Zeiger ein Wort nach links bzw. in die vorherige Zeile der Tabelle. Danach wird (3) bis (7) erneut durchlaufen.*

4.2.1.8 Englisches Verb mit seinen weiteren Elementen

Englische „Phrasal Verben" bilden ihre Bedeutungen mit weiteren Elementen. Das Verb „take" beispielweise kann „nehmen" oder „führen" usw. bedeuten, während „take" mit dem weiteren Element „back", nämlich „take back" bzw. „take sth. back", „zurückziehen" bzw. „etw. zurückgeben" bedeutet. Der folgende Algorithmus sorgt dafür, dass das Verb und seine weiteren Elemente im Ganzen erkannt werden, damit die semantische Bedeutung nicht verloren geht. Durch den TreeTagger wird das weitere Element des Verbs mit der Markierung „RP" gekennzeichnet.

Originalwort	Tagger	Stammform
…	…	…
We	PP	we
also	RB	also
need	VBP	need
to	TO	to
follow	VB	follow
this	DT	this
up	RP	up
and	CC	and
make	VB	make
sure	JJ	sure

Tabelle 6 Beispiel des Englischen Verbs „follow“ mit seinen weiteren Elementen „up“ im Sinne „follow up“.

Voraussetzung

Wenn die Markierung „RP“ gefunden wird, wird die aktuelle Adresse behalten.

1. *Die Variable „Partikel“ wird mit dem aktuellen Wort gesetzt, das mit „RP“ markiert wurde. Der Zeiger geht nach links bzw. ein Wort nach vorne.*
2. *Falls die Markierung mit „V“ anfängt, wird zunächst geprüft, ob das aktuelle Verb mit der gesetzten Variable „Partikel“ in der Tabelle des Phrasal-Verbs gefunden werden kann, sonst geht der Zeiger weiter nach links und (2) wird wiederholt. Falls es sich um ein Phrasal-Verb handelt, wird „Partikel“ mit dem aktuellen Verb verbunden und der Algorithmus terminiert. Wenn das Wort mit der Markierung „RP“ nicht identisch ist mit dem Wert von „Partikel“ dann fragt der Algorithmus, ob dieses Wort gelöscht werden soll.*
3. *Falls die Markierung nicht mit „V“ anfängt, wird der Zeiger ein Wort nach links bzw. eine Zeile nach vorne verschoben und der Prozess kehrt zu (2) zurück.*

4.2.2 Transferwörter

4.2.2.1 Gewählte Wörter

Die Wörter aus der ContextMap werden nur vom Nutzer manuell ausgewählt. Nur die ausgewählten Begriffe werden mit den Suchwörtern zugesetzt.

4.2.2.2 Zentrale aller Wortgruppen

Die Zentroid jeweiliger Wortgruppe wird berechnet. Das nahste Wort wird für seine Gruppe repräsentiert. Die repräsentierten Wörter jeder Gruppe werden in andere Sprache gemeinsam mit den Suchwörtern übertragen. Dies kann voll automatisch durchgeführt werden.

4.2.2.3 Attribute der relevanten Dokumente

Die von relevanten bewerten Dokumenten erzeugte Attribute werden zusammen mit den vom Nutzer gewählten Attributen und den Suchwörtern in die andere Sprache übertragen. Dies wird mit dem Nutzer- bzw. Pseudo-Relevantfeedback erreicht.

4.2.3 Transfermatrix

Die Idee der Transfermatrix kommt aus einer Forschung von Reinhard Rapp. Sein Bericht [RAPP99] bestätigt, dass durch die Übertragung eines Kookkurrenzmusters eines Wortes in deutschen/englischen unverwandten Korpora die Übersetzung zu ca.72% richtig ermittelt wurde. Seine Methode passt genau mit der SENTRAX, weil sie auf dem Wortkookkurrenz basiert. Das elektronische lesbare Wörterbuch wird auch benutzt, um die Assoziationsmatrix zwischen den Wörtern aus den Korpora und den Wörtern aus dem Wörterbuch aufzubauen.

Die Assoziationsmatrizen[23], die den Zusammenhang zwischen den Wörtern in den Korpora und den Wörtern im Wörterbuch beschreiben, werden sowohl für deutsche als auch für englische Texte gebildet. Beim Rapp wird die Kookkurrenzhäufigkeit der mit dem Ausgangwort assoziierten Wörter gezählt und in einem Vektor mit Größe 6 abgelegt. Drei für vor und drei für hinter dem Ausgangwort. Die Assoziationswerte werden im Vektormuster durch die Log-Likelihood Funktion umgewandelt. Die Ähnlichkeit des Musters wird berechnet, um die best mögliche Übersetzung zu treffen.

Weil die indirekte Assoziationsmatrix von SENTRAX die Assoziationen der Wörter in der Sammlung repräsentiert, kann man diese Assoziationsmatrix weiterverarbeiten, indem die Wörter auf den Spalten, die im Wörterbuch nicht gefunden werden, in der Matrix gelöscht werden. Obwohl [RAPP99] die Assoziationszählung und den Assoziationsausdruck für die Wortfolge verwendet hat, wird es für die SENTRAX ohne Wortfolge benutzt, weil die Assoziation bei der SENTRAX ohne Wortfolge aufgebaut wird. Es verbleibt nun nur die im Wörterbuch stehenden Wörter auf den Spalten und die aus der Sammlung gefundenen Wörter auf den Zeilen. Dies wird für beide Sprachen wie beim [RAPP99] durchgeführt. Nach dieser Bearbeitung sind zwei sprachliche wörterbuchbasierte Assoziationsmatrizen entstanden, eine für Deutsch und eine für Englisch. Die Spalten der englischen wörterbuchbasierten Assoziationsmatrix müssen gemäß die Übersetzung des deutschen Wortes angeordnet werden. Wenn mehr als eine Übersetzung entsteht, wird nur das erste übersetzte Wort behalten.

Sei $\mathbf{A_{m,r}}$ die wörterbuchbasierte Assoziationsmatrix der ersten Sprache und

$\mathbf{B_{n,r}}$ die eingeordnete wörterbuchbasierte Assoziationsmatrix der zweiten Sprache entsprechend der ersten Sprache ist,

[23] Die Assoziationsmatrix zwischen den Wörtern in der Korpora und den Wörtern im Wörterbuch

dann $\mathbf{C_{m,n}} = \mathbf{A\Theta B^T}$ ist die Transfermatrix, wobei $\mathbf{C} = \left[c_{ij}\right]_{m,n} = \left[\sum_1^r \left|a_{ir} - b_{rj}\right|\right]_{m,n}$ ist.

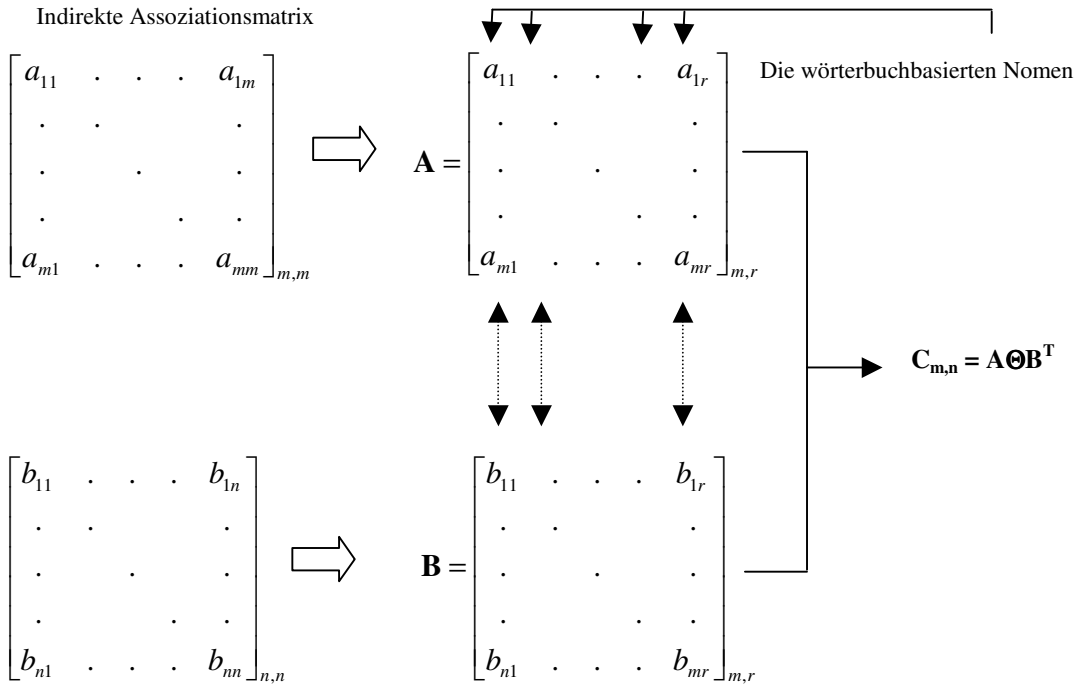


Abbildung 8 Transfermatrix konstruieren

Die Transfermatrix $\mathbf{C_{m,n}} = \mathbf{A\Theta B^T}$ ist die zwischensprachliche Matrix, wobei ihre Zeilen die Wörter in der ersten Sprache und ihre Spalten die Wörter in der zweiten Sprache aus der Sammlung sind. Der Operator **Θ** wird folgendermaßen definiert;

$$\mathbf{C_{m,n}} = \mathbf{A\Theta B^T} = \left[c_{ij}\right]_{m,n}$$

$$c_{ij} = \sum_1^r \left|a_{ik} - b_{ik}\right|$$

Die Komponente c_{ij} kann auch anders definiert werden (siehe Abschnitt 4.2.5.6).

Bei der Übersetzung wird die Zeile des Ausgangwortes in der direkten Assoziationsmatrix zuerst betrachtet. Die *p*-höchsten Werte bzw. Kookkurrenzhäufigkeiten werden markiert. Die entsprechenden Wörter der *p*-höchsten Werte in der Transfermatrix **C** werden gekennzeichnet. Nun werden *p* Vektoren summiert. Die Positionen von den *p* minimalen Werten des summierten Vektors werden behalten. Die Wörter entsprechend allen *p*

Spalten der behaltenen Positionen auf der direkten Assoziationsmatrix der Zielsprache werden aktiviert. Sie werden als Spaltevektor betrachtet. Diese *p* Spaltevektoren werden summiert. Das Wort auf der Zeile mit dem maximalen Wert ist die Übersetzung. Diese Übersetzungsmethode wird „Kookkurrenzübersetzung“ genannt.

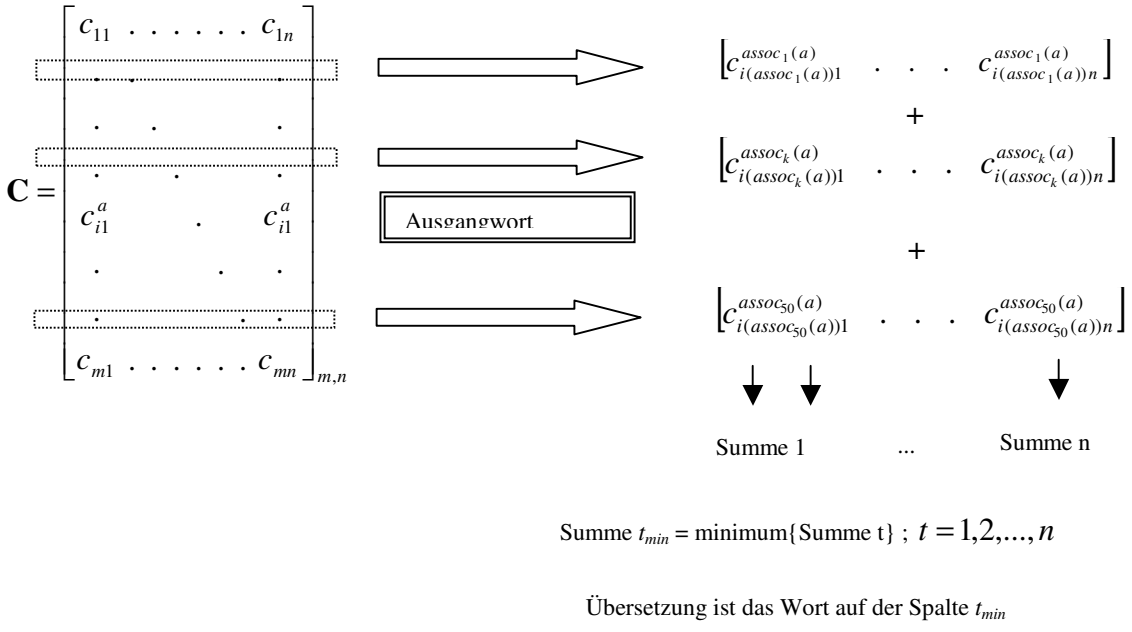


Abbildung 9 Der Übersetzungsprozess

4.2.4 Ähnlichkeit der indirekten Assoziationen

Weil die SENTRAX die Wortassoziation verwendet, um die Wortumgebung auf dem Bildschirm darzustellen, kann die Wortliste der Assoziationshäufigkeit zwischen der Ausgangsprache und Zielsprache verglichen werden. Nach der Übersetzung der übertragenden Attribute geht die SENTRAX zur Zielsprache über, um die indirekte Assoziationslist zu erzeugen. Es gibt wahrscheinlich nicht nur eine Menge von Vertreterwörtern, sondern beliebige endliche Mengen. Nur die beste Liste wird nach dem Vergleich erreicht, weil sie mindestens einen Vertreter von dem Konzeptnetz der ursprünglichen Sprache enthält.

Es wird angenommen, dass es eine Wortliste in der Ausgangsprache und eine in der Zielsprache gibt. Dieses Ähnlichkeitsmaß stützt sich auf das Vektor-Skalarprodukt. Die

Wörter in der Liste werden nach der Übersetzung eingeordnet. Das Gewicht der Wörter, die keine Übersetzung haben, wird null gesetzt. Das Gewicht der entsprechenden Wörter in beider Sprache wird durch den indirekten Assoziationswert übergenommen. Nur die Gewichte werden in Reihefolge auf einem Vektor geschrieben.

Rang	Sim(Suchwort$_{De}$, a_i)	Begriff a_{De}	Sim(Suchwort$_{Eng}$, b_i)	Begriff b_{Eng}
1	1,00	Suchwort$_{De}$	1,00	Suchwort$_{Eng}$
2	0,58	$W_{De,1}$	0,61	$W_{Eng,1}$
3	0,44	$W_{De,2}$	0,50	$W_{Eng,2}$
4	0,37	$W_{De,3}$	0,35	$W_{Eng,3}$
5	0,25	$W_{De,4}$	0,23	$W_{Eng,4}$
6	0,22	$W_{De,5}$	0,21	$W_{Eng,5}$
7	0,13	$W_{De,6}$	0,13	$W_{Eng,6}$
8	0,11	$W_{De,7}$	0,12	$W_{Eng,7}$
9	0,10	$W_{De,8}$	0,12	$W_{Eng,8}$
10	0	$W_{De,9}$	0	$W_{Eng,9}$

Tabelle 7 Die Assoziationshäufigkeiten entsprechend dem Suchwort werden nach der Verstärkung eingeordnet. $W_{De,i}$ hat die Übersetzung $W_{Eng,i}$, i = 1,2,...,9. Ab i = 10 findet sich kein Übersetzungspaar.

Von der obigen Tabelle können zwei Vektoren formuliert werden, und zwar
$\overrightarrow{W}_{De} = (1.00, 0.58, 0.44, 0.37, 0.25, 0.22, 0.13, 0.11, 0.10, 0)$ und
$\overrightarrow{W}_{Eng} = (1.00, 0.61, 0.50, 0.35, 0.23, 0.21, 0.13, 0.12, 0.12, 0)$. Die Ähnlichkeit zwischen den Vektoren wird durch das Skalarprodukt ermittelt,
$\ddot{A}hnlichkeit(\overrightarrow{W}_{De}, \overrightarrow{W}_{Eng}) = \overrightarrow{W}_{De} \cdot \overrightarrow{W}_{Eng}$. Durch die Ähnlichkeit der Assoziationslisten kann einschätzt werden, welche vertretende Wortgruppe die Zieldokumente korrekt beschreiben könnte.

4.2.5 Graphabgleichung

Um in der ContextMap-Funktion nach den höher kookkurrierenden Wörtern in Cluster aufteilen zu können, muss die Distanzdifferenz (Unähnlichkeit der Distanz) erst berechnet werden. In diesem Schritt wird ein hierarchischer Baum erzeugt, wobei ein Blatt das Wort und eine Kante den Abstand repräsentiert. Dieser Prozess findet sich nach dem

Clusterverfahren in der ContextMap-Funktion. Die Vorgängerknoten sind nur logische Knoten und beinhalten keinen Begriff, während jedes Blatt des Baums einen Begriff besitzt. Die Vorgängerknoten verknüpfen die Begriffsblätter, wonach die Begriffe mittels Schwellwert gruppiert werden können. Der Algorithmus (siehe Abschnitt 4.2.5.2) leistet eine Umformung des Baums von der ContextMap-Funktion zum Graphen der Begriffsknoten. Der induzierte Untergraph von zwei Graphen wird aufgebaut (siehe Abschnitt 4.2.5.3), damit man die Ähnlichkeit der Strukturen berechnen kann. Hier kann die Graphabgleichung verwendet werden, um die zwei Strukturen der zu betrachtenden Wörter zu vergleichen. Die Basis für die Berechnung der Strukturähnlichkeit besteht aus der Ähnlichkeit des Konzepts und der Ähnlichkeit der Relation.

4.2.5.1 Einige grundlegende Begriffe über Graphen[24]

- Ein Graph G ist eine endliche nichtleere Menge $V(G)$ von Knoten und eine (möglicherweise leere) Menge $E(G)$ von Kanten, wobei eine Kante eine Zwei-Element-Mengen von $V(G)$ ist. Die Menge $V(G)$ wird Knotenmenge des Graphen G genannt und $E(G)$ wird Kantenmenge des Graphen G genannt.

- Für einen Knoten v des Graphen G ist sein Nachbar folgendermaßen definiert:

  $$N(v) = \{u \in V(G) / vu \in E(G)\}$$

  und sein Grad $\deg_G(v)$ ist die Anzahl der zu v adjazenten Knoten, d.h. $\deg_G(v) = |N(v)|$.

- Wenn $e = uv$ eine Kante des Graphen G ist, sagt man, dass e und u bzw. e und v inzident miteinander sind. Wenn e und f die unterschiedliche Kanten und inzident mit demselben Knoten sind, sind e und f die adjazenten Kanten.

- Ein Graph H ist ein Untergraph des Graphen G, wenn $V(H) \subseteq V(G)$ und $E(H) \subseteq E(G)$.

[24] vgl. [CHOE93]

- Ein spezieller Untergraph $G - S$ entsteht, wenn man G alle Knoten, die nicht in S sind, belässt und alle Kanten, die mit einem Knoten aus S inzidieren, entfernt. Dabei muss stets $S \subset V(G)$ sein.
- H heißt ein induzierter Untergraph des Graphen G, falls $E(H) = E(G) \cap C_2^{V(H)}$ gilt, d.h. H enthält alle Kanten zwischen den Knoten in $V(H)$, die auch in G vorhanden sind. Für die Knotenmenge S, $S = V(H) \subset V(G)$, kann man den induzierten Untergraph H als $\langle S \rangle$ bezeichnen, wenn $E(\langle S \rangle) = E(G) \cap C_2^{V(\langle S \rangle)}$ gilt. Der Untergraph $\langle S \rangle$ ist der von S induzierte maximale Untergraph.
- Ein Kantenzug W in einem Graph G besteht aus einer Folge $u_1, u_2, ..., u_n$ von verschiedenen Knoten mit $u_i u_{i+1} \in E(G)$ für alle i. Wenn die u_i paarweise voneinander verschieden sind, so spricht man von einem Weg, bzw. von einem Kreis im geschlossenen Fall.
- Seien u und v Knoten in einem Graph G. Man sagt, u ist verbunden mit v, wenn G einen Weg von *u* nach *v* enthält.
- Der Graph G ist zusammenhängend, wenn ein Knoten u mit v für alle Knoten u, v in $V(G)$ verbindbar ist.

4.2.5.2 Algorithmus (Wörterbaum zum Graph)

Voraussetzung: Wörterbaum B durch die ContextMap-Funktion.

Gegeben seien u, v die Blätter des Wörterbaum,

$L(B)$ die Menge der Blätter des Wörterbaums,

z_i die Vorgängerknoten und

w_j das Gewicht bzw. die Distanz zwischen den Knoten.

1. *$\forall u \in L(B),\ v \in L(B) \wedge v \neq u$, wird der kürzeste Weg von u nach v gesucht.*
2. *Der Weg $uz_i v$ repräsentiert die Beziehungsdistanz zwischen dem Begriff u und dem Begriff v durch die Vorgängerknoten des Baums. Ein neuer Graph wird*

hier erzeugt, indem die Vorgängerknoten des Weges z_i weggenommen werden und die benachbarte Kanten vereinigt werden. Während der Vereinigung der Kanten werden die Distanz w_j aufsummiert.

3. *Die Distanz zwischen dem Begriff u und dem Begriff v in neuem Graph H wird auf $\frac{1}{2}\sum_j w_j$ gesetzt, wobei w_j die Distanz zwischen den Knoten in dem kürzesten Weg uz_iv ist. Der abgeleitete Graph H wird hier „Konzeptsgraph" genannt.*

4.2.5.3 Induzierter Untergraph von zwei Graphen

Satz 1 Gegeben seien G_1 und G_2 die zusammenhängenden (gewichten) Graphen, deren Knoten gekennzeichnet wurden. Gegeben sei S die Menge der Knoten, wobei $S = V(G_1) \cap V(G_2)$ ist, und X die Menge der Kanten, wobei $X = E(G_1) \cap E(G_2)$ ist. Der induzierte Untergraph von S, sogenannt $\langle S_X \rangle$, kann erbaut werden, indem seine Kantemenge $E(\langle S_X \rangle)$ die Untermenge der Kantemenge X, $E(\langle S_X \rangle) \subseteq X$. Falls der Grad einen Knoten in $\langle S_X \rangle$ null ist, wird dieser Knoten weggelassen ($\forall u / |N(u)| = 0$ wird $\langle S_X \rangle - \{u\}$ durchgeführt). Der ergebene Graph $H = \langle S_X \rangle - \{u / |N(u)| = 0\}$ ist der maximale zusammenhängende Untergraph von G_1 und G_2.

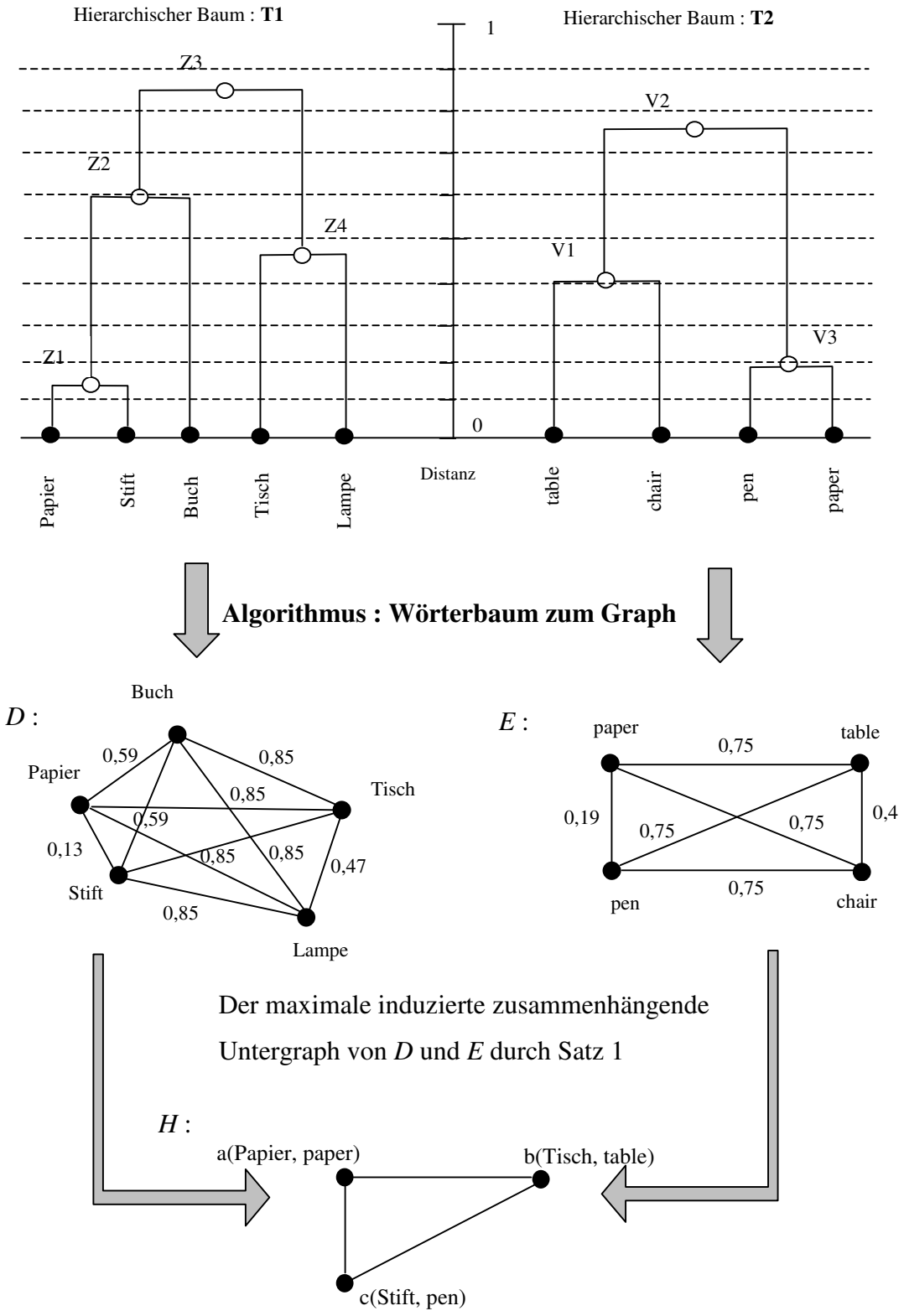


Abbildung 10 Der Prozessablauf zur Vorbereitung der Graphabgleichung. Die Distanzachse sagt uns, wie weit die Wörter vom Zentroid einer beliebigen Gruppe entfernt ist.

Die Stammidee der Graphabgleichung kommt aus der Arbeit von Montes-y-Gómez et al. Die Definitionen in Abschnitt 4.2.5.4-4.2.5.6 werden von [MLG00] zum Ähnlichkeitsmaß der Graphen adaptiert.

4.2.5.4 Ähnlichkeit des Konzepts

Definition : graphrepräsentierendes Konzept

Für den Graph *G* repräsentiert sein Knoten ein Konzept.

Definition : Konzeptähnlichkeit

Gegebene zwei Graphen, nämlich G_1 und G_2. Der induzierte Untergraph G_c wird durch den Satz 1 (Abschnitt 4.2.5.3) erfolgt, wobei seine Knoten die Übersetzung entsprechen und seine Kanten die Relation repräsentieren. Die Konzeptähnlichkeit den beiden Graphen lautet:

$$S_c = \frac{2n(G_c)}{n(G_1)+n(G_2)}$$

wobei $n(G_i)$ die Anzahl des Knotens in G_i ist.

4.2.5.5 Ähnlichkeit der Relation

Definition : Relationen der Konzepte

Für den Graph *G* repräsentiert seine Kante eine Relation.

Definition : Relationsähnlichkeit

Gegebene zwei Graphen, nämlich G_1 und G_2. Der induzierte Untergraph G_c wird durch den Satz 1 (Abschnitt 4.2.5.3) erfolgt, wobei seine Knoten die Übersetzung entsprechen und seine Kanten die Relation repräsentieren. Die Relationsähnlichkeit den beiden Graphen lautet:

$$S_r = \frac{2e(G_c)}{e(G_1)+e(G_2)}$$

wobei $e(G_i)$ die Anzahl der Kante in G_i ist.

4.2.5.6 Ähnlichkeit der konzeptionellen Graphabgleichung

Definition : gesamte konzeptionelle Graphähnlichkeit

Gegebene zwei Graphen, nämlich G_1 und G_2. Der induzierte Untergraph G_c wird durch den Satz 1 (Abschnitt 4.2.5.3) erfolgt, wobei seine Knoten die Übersetzung entsprechen und seine Kanten die Relation repräsentieren. Die gesamte konzeptionelle Graphähnlichkeit zwischen G_1 und G_2 wird folgendermaßen definiert:

$$S = S_c \times (a + b \cdot S_r)$$

wobei $a = \dfrac{2n(G_c)}{2n(G_c) + e_{G_c}(G_1) + e_{G_c}(G_2)}$

$e_A(B)$ die Anzahl der Kante in dem induzierten Untergraph A des Graphen B

und $b = 1 - a$ ist.

Die Ähnlichkeitsmaß kann in anderer Form vom Ähnlichkeitskoeffizient umwandelt werden. Es kommt darauf an, wie der Graph aussieht. Die bekannten Ähnlichkeitskoeffizienten sind folgendermaßen:

Dice's Koeffizient $\dfrac{2 \cdot |X \cap Y|}{|X| + |Y|}$

Jaccard's Koeffizient $\dfrac{|X \cap Y|}{|X \cup Y|}$

Kosinus Koeffizient $\dfrac{|X \cap Y|}{|X|^{1/2} + |Y|^{1/2}}$

Bei Definition im obigen Abschnitt 4.2.5.4 und 4.2.5.5 wird die Dice's Koeffizient verwendet. Sie passen dem Graph ohne Gewicht sehr gut. Aber die umgeformten Graphen

vom Wörterbaum haben das Gewicht, nämlich Unähnlichkeitsdistanz. Dadurch werden die Definitionen zum gewichten Graph adaptiert.

4.2.5.7 Graphabgleichung mit dem gewichten Graph anpassen

Da die Konzeptsgraphen, die von den hierarchischen Bäumen abgeleitet werden, sowohl die Relation zwischen zwei Knoten auch als die Distanz des Clusters behält, passt die Ähnlichkeit der konzeptionellen Graphabgleichung mit der Dice's Koeffizient den Konzeptsgraphen nicht direkt. Die Distanz ist eine wichtige Information. Sie beschreibt, wie stark der Zusammenhang zwischen den Wörtern bzw. den Attributen ist. Bei dem Konzeptsgraph wird daher der Kosinus-Koeffizient statt dem Dice's Koeffizient verwendet, damit die Ähnlichkeit der gewichteten Graphen nicht verloren wird.

Definition : Konzeptähnlichkeit den Konzeptsgraphen

Gegebene zwei Konzeptsgraphen K_1 und K_2. Der verbundene Untergraph K' wird von den Konzeptsgraphen K_1 und K_2 ordnungsgemäß des Satzes 1 (Abschnitt 4.2.5.3) induziert. Die Konzeptähnlichkeit den beiden Konzeptsgraphen lautet:

$$S_c = \frac{n(K')}{n(K_1)^{1/2} + n(K_2)^{1/2}}$$

wobei $n(K_i)$ die Anzahl des Knotens in K_i ist.

Definition : Relationsähnlichkeit den Konzeptsgraphen

Gegebene zwei Konzeptsgraphen K_1 und K_2. Der verbundene Untergraph K' wird von den Konzeptsgraphen K_1 und K_2 ordnungsgemäß des Satzes 1 (Abschnitt 4.2.5.3) induziert. Die von K' abgeleitete Untergraphen K_1' und K_2' haben die Gewichte wie K_1 bzw. K_2. Die Relationsähnlichkeit den beiden Konzeptsgraphen lautet:

$$S_r = \frac{\frac{1}{2}\vec{e}(K_1') \cdot \vec{e}(K_2')}{\|\vec{e}(K_1)\|_2 + \|\vec{e}(K_2)\|_2}$$

wobei $\vec{e}(K_i)$ und $\vec{e}(K_i')$der Relationsvektor zwischen eingeordneten Wörtern in K_i bzw. K_i',

$\|\cdot\|_2$ die L_2-Norm ist.

Obwohl bei diesen gewichten Konzeptsgraphen die Relationsähnlichkeit S_r niemals Null wird, kann die gesamte konzeptionelle Graphähnlichkeit S wie die Definition in Abschnitt 4.2.5.6 verwendet werden. Nur einziger Ausdruck muss geändert werden, nämlich a. Um den Ausdruck a mit der Konzeptsgraphähnlichkeit anzupassen, wird a folgendermaßen definiert:

$$a = \frac{n\left(K_1'\right)}{n\left(K_1'\right) + \left\|\vec{e}\left(K_1'\right)\right\|_2 + \left\|\vec{e}\left(K_2'\right)\right\|_2}$$

4.3 Monolinguales Modell

Bevor eine bilinguale Suche mit der SENTRAX erfolgen kann, erscheint es sinnvoll, den Prozess im Monolingualen zu betrachten, weil die Sprachen oft sehr unterschiedliche Strukturen haben und diese erst angeglichen werden müssten. Zuerst muss die SENTRAX für den deutschen und englischen Korpus angepasst werden. Die Texte

werden zunächst durch die Tagger-Anwendung, TreeTagger, bearbeitet. Nur die benötigten Wortarten (siehe Abschnitt 4.2.1.2) verbleiben im jeweiligen Dokument in drei positionsparallelen Dateien. Die vorgenannten Funktionen sind mit TIHO automatisch durchführbar (siehe Abschnitt 7.3). Die Vorverarbeitung muss in dem Verfahren eingebaut werden, um die Gemeinsamkeiten der Sprachen herauszuarbeiten, je nach der Sprache. Die Verknüpfung der reduzierten Formen zu den ursprünglichen Texten wird gebildet, damit der Rückgriff auf die Varianten im Text durch die Stammform möglich ist. Danach lassen sich die direkte und indirekte Assoziationen durch die SENTRAX erzeugen.

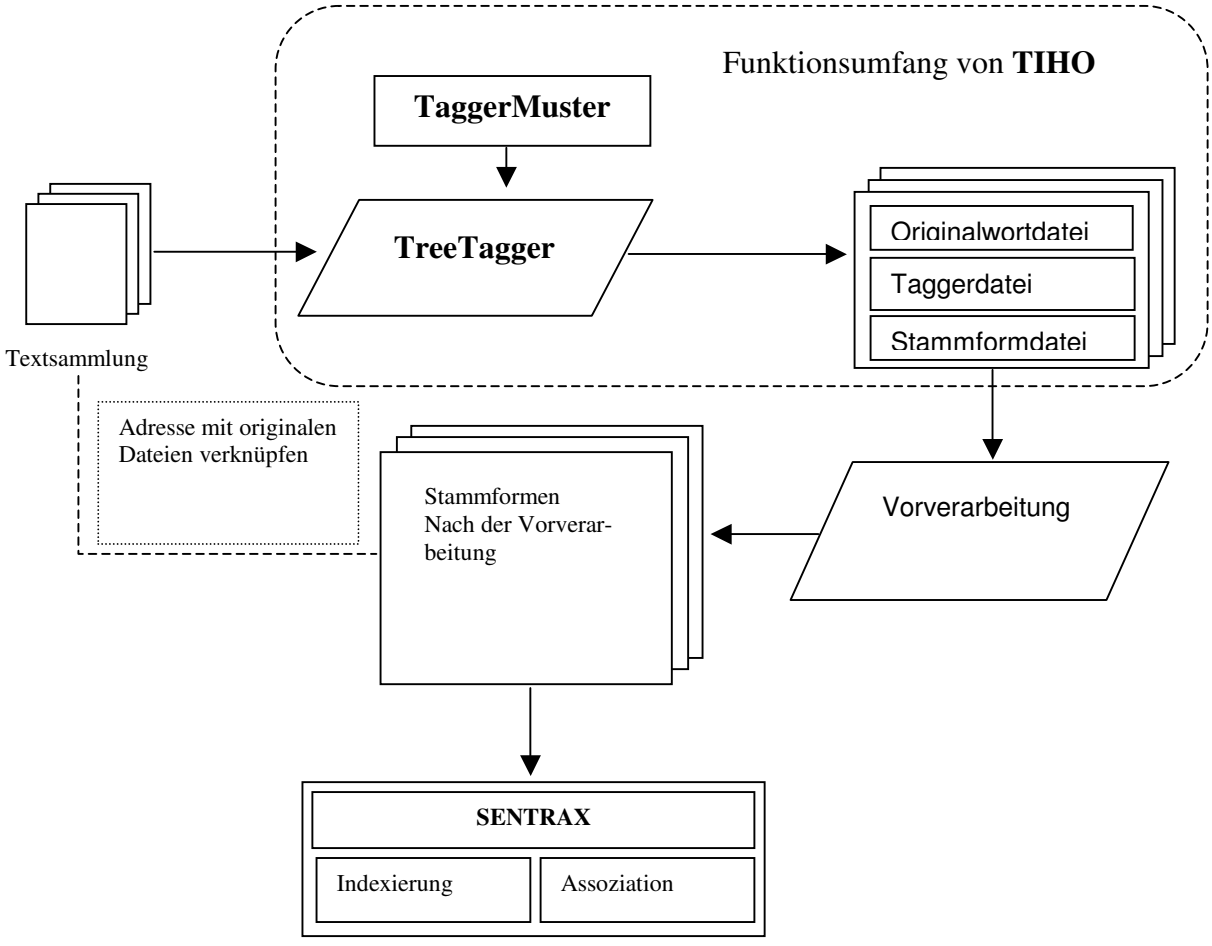


Abbildung 11 Monolinguales Modell mit der Vorverarbeitung für SENTRAX

4.3.1 Monolinguales Modell für deutsche Sprache

Folglich der Abbildung 11 soll die Vorverarbeitungen für deutsche Sprache in den Ablauf eingebaut werden. Die Anerkennungsprozesse für das deutsche Kompositum und die deutsche Mehrwortgruppe sowie das trennbare Verb (siehe Abschnitt 4.2.1) werden in der D-SENTRAX, deutsche SENTRAX, integriert.

4.3.2 Monolinguales Modell für englische Sprache

Die Vorverarbeitung für die englische Sprache muss die Anerkennung der Mehrwortgruppe und des englischen Verbs mit seinen weiteren Elementen (siehe Abschnitt 4.2.1) beinhalten. Sie werden an der Stelle nach dem Markierungsprozess durchgeführt. Die endgültige SENTRAX wird E-SENTRAX, englische SENTRAX, genannt.

4.4 Bilinguales Modell

4.4.1 Überblick

Nach der Erzeugung des deutschen und englischen SENTRAX-Containers werden die beiden monolingualen Systeme durch eine sprachliche Brücke verbunden, in diesem Fall ist die Brücke das elektronische lesbare Wörterbuch bzw. Transfermatrix. Die Auswahlmethode der übertragenden Wörter zu einer Anfrage in Zielsprache wird betrachtet, um das best mögliche Suchergebnis zu erhalten. Im Fall der parallelen Korpora ist das denkbar beste Ergebnis das übersetzte Paar. Obwohl unser Modell nur eine bilinguale Suche zwischen der deutschen und englischen parallelen Textsammlung darstellt, kann man aber auf anderen Sprachpaaren, z.B. englisch-thailändisch, analog verfahren. Außerdem kann die englische Sprache als Umsetzungssprache (siehe Abschnitt 3.1.2.6) für bilinguale Systeme, z.B. deutsch-thailändisch, benutzt werden. Mit der Konzeptübertragung würde die bilinguale bzw. multilinguale Suche mehr toleranter sein als dies mit einfachen Wortübersetzungen möglich wäre.

In der Realität gibt es zufällig große Mengen von Texten, die dasselbe Thema in verschiedenen Sprachen beschreiben, z.B. Zeitungen, Berichte, Bücher, Webseiten usw. Solche Texte werden dann *vergleichbare Texte* genannt. Unser bilinguales Suchmodell wird auch bezüglich der vergleichbaren (nicht-parallelen) Texte daraus überprüft, ob die Übertragung durch das Konzeptnetz weiterhin gut funktioniert.

4.4.2 Brücke zwischen zwei Sprachen

Die Brücke zwischen den monolingualen IR-Suchsystemen ist der wichtigste Teil des krosslingualen Systems. Das elektronische lesbare Wörterbuch wird an dieser Stelle am meisten verwendet. Fehlende Fachwörter sind dabei ein großes Problem, welches die gesamte Leistung der bilingualen Suche deutlich verringert. Nicht nur die Fachwörter, sondern auch die Mehrdeutigkeit, die wegen vieler Übersetzungsmöglichkeiten entstehen, sollten vermieden werden. Die Methode von [RAPP99] wird in unserem Modell angepasst, weil sie auf der Wortkookkurrenz wie die SENTRAX basiert. Von [RAPP99] wird berichtet, dass seine Übersetzungsmethode auf den unverwandten englischen und deutschen Korpora gut funktioniert.

Die Idee zur Transfermatrix stammt aus der Methode von Rapp [RAPP99], die Transfermatrixbildung erfolgt bei uns auf andere Weise. Die Kookkurrenzhäufigkeit von Rapp wird erst in der Vektorform je nach der vorkommenden Position gezählt und danach durch ein gewichtetes Muster summiert (vgl. [RAPP99]). Bei der SENTRAX dagegen erhält man die Kookkurrenzhäufigkeit durch die direkte und indirekte Assoziation. Das normale elektronisch lesbare Wörterbuch steht zur Verfügung, um es entweder direkt als Brücke oder für die Bildung einer Transfermatrix zu nutzen.

4.4.3 Parallele Korpora

Der hierfür genommene Korpus vom Europarl-Projekt enthält englisch-deutsche parallele Dateien vom Jahr 1996 bis zum Jahr 2003. Bei diesen parallelen Korpora handelt es sich um Sitzungsberichte von Mitgliedern des Europäischen Parlaments. Die Größe der

originalen Dateien ist insgesamt für die deutsche Sprache 174 Mb bzw. für die englische Sprache 155 Mb. Jeweils aufgeteilt nach dem Sitzungstag ergeben sich 488 Teile. Nach der neuen Verteilung durch die Markierung „Chapter ID“ ist die Anzahl der Texte auf Deutsch 4.048 bzw. auf Englisch 3.989, weil die Anzahl der Markierung „Chapter ID“ in den deutschen und englischen Texten nicht gleich ist.

4.4.4 Struktur des Modells

Nachdem zwei Container der monolingualen SENTRAX vorbereitet wurden, kann der Nutzer die bilinguale SENTRAX in einer Sprache starten. Die Anfrage wird in der Ausgangsprache eingegeben. Die Tippfehler und Schreibvarianten werden durch die LexicoMap entdeckt. Der Nutzer kann natürlich seine Fehler korrigieren sowie andere Schreibvarianten dazu eingeben. Die ContextMap wird mit seiner Anfrage und zusätzlichen Schreibweise erzeugt. Die häufigen zusammenauftretenden Wörter werden in die Gruppe auf dem Bildschirm eingebettet, damit der Nutzer andere verwandte Wörter als Attribute frei wählen kann, um sein Suchkonzept zu beschreiben.

Die ausgewählten Attribute und ihre Relationen werden gemerkt, damit ihre Strukturähnlichkeit mit dem Konzeptnetz in der Zielsprache später berechnet werden kann. Die ausgewählten Attribute werden entweder mit dem elektronischen Wörterbuch oder mit der Transfermatrix in die Zielsprache als Anfrage übertragen (an der Stelle 2 in der Abbildung 12). Die übertragenen Attribute hängen davon ab, welche Auswahlmethode verwendet wird (siehe Abschnitt 4.2.2). Jede mögliche Kombination der Übersetzungen bildet ein eigenes Konzeptnetz. Alle Konzeptnetze werden mit der gemerkten Konzeptstruktur der ursprünglichen Sprache verglichen, um die ähnlichste Struktur in der Zielsprache zu finden. Entweder die Ähnlichkeit der indirekten Assoziationen (im Abschnitt 4.2.4) oder die Graphabgleichung (im Abschnitt 4.2.5) ist hier für die Entscheidung verantwortlich. Der beste Vertreter ist die Gruppe mit den übersetzten Attributen, wobei ihre Struktur mit dem ursprünglichen Konzept am vergleichbarsten ist. Der beste Vertreter sollte in der Zielsprache die beste Trefferliste der Dokumente sowie die Liste der ähnlichen Dokumente durch die SimilarDoc-Funktion erzeugen.

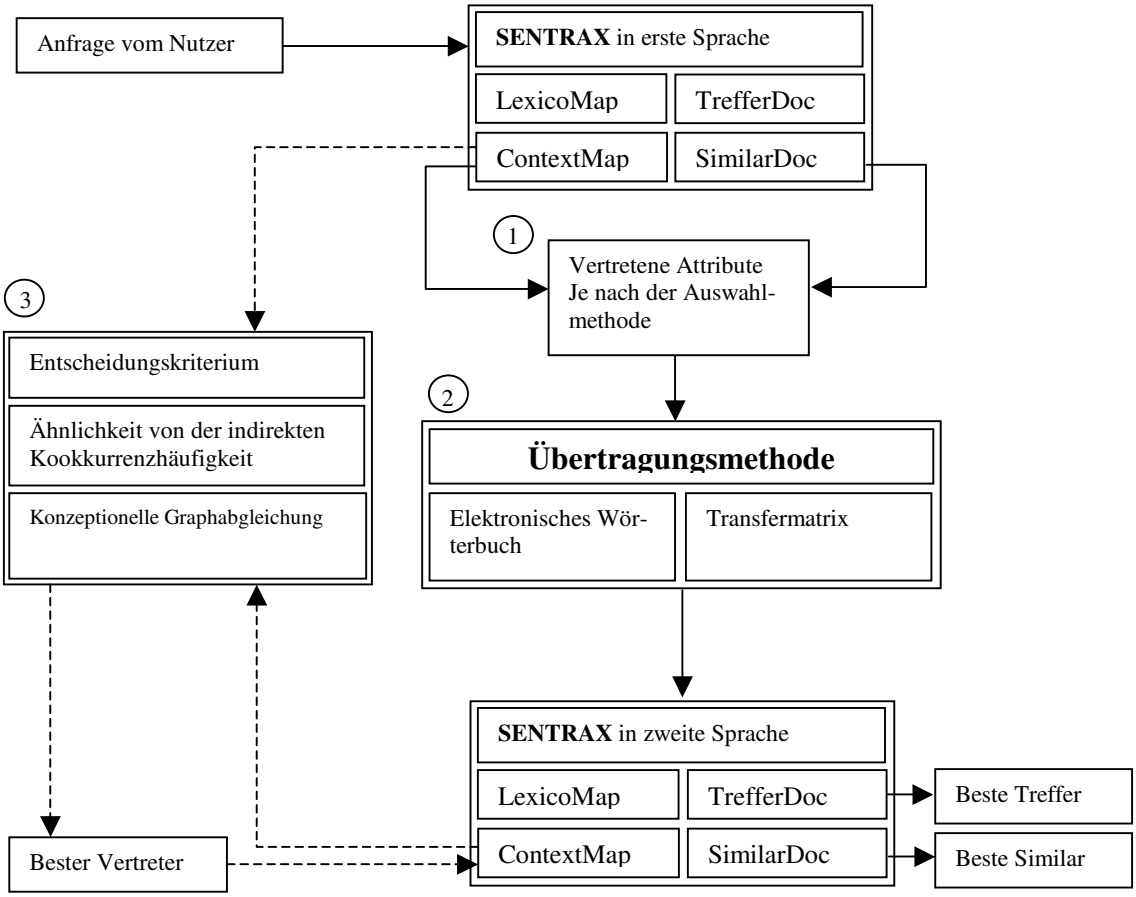


Abbildung 12 Der Datenfluss der bilingualen Suche durch die SENTRAX

An der Stelle 1 in der Abbildung 12 lässt sich untersuchen, welche im Abschnitt 4.2.2 beschriebene Auswahlmethode mit dem Modell am besten passt. Die Übertragungsmethode an der Stelle 2 wird unter dem Gesichtpunkt gewählt, ob die Transfermatrix oder das elektronische Wörterbuch besser sind zur Bildung des Konzeptnetzes. An der Stelle 3 werden zwei Strukturen des Konzeptnetzes verglichen, die aus der Ausgangsprache und aus der Zielsprache stammen. Zwei Vergleichsmaße, nämlich die Ähnlichkeit der indirekten Kookkurrenzhäufigkeit und die konzeptionelle Graphabgleichung (siehe Abschnitt 4.2.4 bzw. 4.2.5), werden darauf geprüft, welches für die Konzeptstruktur zum Vermessen sinnvoller ist.

4.5 Hypothese

- Wir gehen aus von zwei Datensammlungen "D" und "E", die parallel seien. (Eine durch die Parallelität bereits mitgegebene Zuordnung dient lediglich zur späteren Überprüfung unserer Entscheidung, ob das von der Maschine mittels des neuen, schon skizzierten Vorgehens gefundene Dokument das gesuchte Zieldokument in der anderen Sprache ist.) Für D und E werden zunächst unabhängig die SENTRAX-Indexe erzeugt.

- Die Vermutung ist, dass bei dieser Datenlage die beiden internen Konzeptnetze eine ähnliche Struktur haben. "Ähnlich" im Sinne der parallelen Dokumentenpaare: Die Umgebung eines Dokuments in seinem Index "entspricht" der Umgebung seiner Übersetzung im anderen Index. Sollte diese Vorstellung zutreffen, dann müsste die (automatische) Übertragung der einem Dokument hier zugeordneten Wortgruppen ein Cluster von Wörtern dort erzeugen, zu denen das Zieldokument unter allen am besten passt. Als Vorteil bei dieser Methode ist zu erwarten, dass Mehrdeutigkeiten durch die (später vollautomatische) Übersetzung nicht stören, da Bestandteile, die keine Korrespondenzen in den Dokumenten haben, durch den SENTRAX Automatismus unwirksam bleiben. Hierdurch entsteht eine enorme Reduktion der kombinatorischen Möglichkeiten.

  Der Umstand, der hier ausgenutzt wird, vergleicht sich mit folgender, "natürlicher" Situation: Wenn man mit einem Menschen spricht, der unsere Sprache nicht besonders gut kann, dann versteht er Gesprächspassagen nicht, in denen Wörter oder Phrasen vorkommen, die ihm fremd sind. Er versteht eben nur das, was in seinem Gehirn eine Korrespondenz in seiner Muttersprache besitzt. Insofern bleibt zuweilen eine große Ausdrucksvielfalt auf unserer Seite nutz- und wirkungslos, da das Verstehenspotenzial auf der anderen Seite sehr eingeschränkt ist.

- Die Suchwörter und ihre umgebungsbedingte assoziierte Begriffe, die im "ContextMap" des einzigen 100%-Treffers im Container "D" auftreten, werden dann Schlüsselwörter bzw. vorgeschlagene Zusatzbegriffe für die Suche im Container

"E". Eventuelle Mehrdeutigkeiten im Wörterbuch werden einfach mitgenommen. Die in den Korpora existierende Übersetzung der Schlüsselwörter sollte zum parallelen englischen Dokument entsprechend des vorherigen deutschen Dokuments hinführen.

- Diese Vermutung und Methode ist symmetrisch, lässt sich also auch von "E" nach "D" verwenden.
- Bei einem großen Container, sollte die Durchmischung der Begriffe auf dem Konzeptnetz gemäß derselben Anfrage möglichst gering werden. Sollte dieses geschehen, könnte das korpusbasierte Semantiknetz mit der ContextMap-Funktion erschaffen werden.
- Gäbe es kein paralleles Dokument in der Zielsprache entsprechend der Anfrage, sollte das Konzeptnetz zu den anderen ähnlichsten Dokumenten führen.
- In der Situation, dass der Zielcontainer viel größer oder viel kleiner als der Ausgangcontainer ist, würde die Kookkurrenzhäufigkeit nur gering abweichen, aber die Antwort, nämlich das parallele Dokumentenpaar, sollte noch in dem Dokumententreffer dargestellt werden. Die bilinguale Suche durch das Konzeptnetz könnte noch funktionieren.
- Hätte der Zielcontainer mehrere Sprachen, sollte die bilinguale Suche durch die SENTRAX gut funktionieren, weil der Zusammenhang zwischen Wörtern unterschiedlicher Sprachen nur selten entstehen würde.
- Für die nicht-parallele Textsammlung werden die deutschen und englischen Dokumente durch die SENTRAX unabhängig geführt. Befänden sich die relevanten Dokumente in beiden Containern, sollten die Konzeptnetze miteinander vergleichbar sein. Würde der Zusammenhang zwischen den Konzeptnetzen zu schwach sein, hätten die entsprechenden Dokumente möglicherweise keine Relation zueinander.

5 UNTERSUCHUNGEN UND DISKUSSIONEN

Die im letzten Kapitel formulierten Hypothesen werden in diesem Kapitel diskutiert. Dabei ergeben sich vier Fälle: (1) der Standardfall, (2) Sonderfälle, (3) die Konzeptnetzänderung und (4) Suche im nicht-parallelen Korpus. Die Verhältnisse im Suchprozess werden für jeden der vier Fälle beobachtet. Die Ergebnisse im Standardfall und in den Sonderfällen bestätigen, dass die bilinguale Suche mittels Konzeptnetzen nicht nur das gesamte Such-Konzept bewahren kann, sondern auch stabil ist. Schließlich wird das Verhalten des Konzeptnetzes bei Containeränderung diskutiert. Dabei wird beobachtet, was auf vergleichbaren Containern erzeugt und in der ContextMap angezeigt wird. Die Übersetzung wird stets mit dem Online-Wörterbuch http://www.leo.org/ manuell vorgenommen; in der später kommenden Ausbaustufe soll hier auch eine automatische Übersetzung eingeschaltet werden können. Die Einstellung der Parameter wird sich dabei an der hier vorgestellten "halbautomatischen" Prüfung der Hypothese orientieren.

Der für die Untersuchungen genommene Korpus vom Europarl-Projekt beinhaltet parallele englisch-deutsche Dateien, die sich nach Sitzungstag und Gesprächthema orientieren.

5.1 Standardfall

Beim Standardfall werden die Untersuchungen auf einem parallelen Container durchgeführt. Der Container “D“ bezeichnet die deutsche Textsammlung, “E“ die englische Textsammlung. Die hierfür genommenen parallelen Dateien sind aus dem Jahr 2001 mit einer Größe von 25 MB für D und 22 MB für E. Mit einfachen Suchwörtern startet der Prozess auf dem Index zu D. Zusätzliche Terme werden mittels der grafischen Darstellung der ContextMap gefunden und ausgewählt. Diese kommen zusätzlich in die Anfrage, bis schließlich genau ein (100%-) Trefferdokument erreicht bzw. angezeigt wird. Einige Terme auf der ContextMap (oder eine sehr nahe Auswahl davon), die aus dieser Anfrage herrühren und aus dem Index mit aufgerufen wurden, werden dann zunächst

manuell mit Hilfe des Wörterbuches übersetzt. Die Übersetzungen der auf der Ausgangsseite ausgewählten Begriffe werden nun sämtlich als Anfrage in den Zielindex gegeben. Das am besten getroffene Dokument (ggf. mehrere) wird nun mit dem Dokument aus der Anfangssprache verglichen. Zusätzlich werden auch die ähnlichen Dokumente, die sich mittels der SimilarDoc-Funktion (in jedem der beiden Indexe unabhängig voneinander) ergeben, mit dem Ausgangsdokument verglichen.

Bezüglich der Hypothese, dass die bilinguale Suche durch das Konzeptnetz für den parallelen Korpus funktionieren sollte, ist hauptsächlich zu prüfen, ob die SENTRAX mit der Kookkurrenztechnik jeweils "das" parallele Dokument finden kann. Bei dieser Untersuchung wird dabei erst die anfängliche Anfrage formuliert. Damit wird dann die ContextMap erzeugt, gegebenenfalls in Schritten. Die Übersetzungen der angezeigten Terme werden mittels Online-Wörterbuch angefertigt und in die englische SENTRAX eingegeben. Solange es mehr als ein Trefferdokument gibt, wird ein weiteres Wort aus der deutschen ContextMap genommen, übersetzt und in dieser Menge eingegeben. Solche Wörter werden somit zur englischen Anfrage hinzugefügt.

5.1.1 Lernen während der Suche

Hier wird die Situation imitiert, dass ein Suchwort keine Übersetzung in dem maschinellen lesbaren Wörterbuch hat. Aus den ermittelten Begriffen könnte man erahnen, zu welchen Themen die Kombination der Begriffe gehören kann. Der Sucher kann dabei „lernen“, ob seine Suchanfrage zum benötigten Thema hinführt oder welche Begriffe noch hinzugefügt werden müssen, um sein Suchkonzept genauer zu beschreiben. Außerdem kann das Problem der Unübersetzbarkeit durch die Auswahl der angebotenen Wörter in der Zielsprache vermindert werden.

Experiment

Das erste Beispiel beginnt mit der Eingabe der Terme *Entwicklungszielen, Bildung, Qualität, Verbesserung* in den D-Index. Diese Anfrage führt bereits zu einem einzigen Dokument. Aber mit ihrer direkten Übersetzung, nämlich mit *education, quality, impro-*

vement, werden mehrere Dokumente auf der E-Seite getroffen. Weil das Online-Wörterbuch keine Übersetzung von *Entwicklungsziel* hat, wird einfach ohne dieses Wort weitergearbeitet. In der D-ContextMap tauchen weitere Wörter, unter anderen *Armut, Entwicklungsländer, Analphabetentum, Zusammenarbeit, Förderung usw.* auf, die zur Umgebung des einen Dokuments auf der D-Seite gehören. Diese Begriffe müssen keineswegs in diesem D-Dokument enthalten sein. Trotzdem bilden sie die bestpassenden Kookkurrenzen. Sie lehren uns zum Beispiel, dass es in der Dokumentensammlung ein Thema der Art „Qualitätsverbesserung und Förderung der Bildung in Entwicklungsländern, wo die Menschen arm sind und nicht lesen können“ gibt. Für dieses Thema eignen sich diese Begriffe zur Übersetzung und Eingabe in die E-SENTRAX. Das einzige parallele englische Dokument taucht jetzt tatsächlich in der E-ContextMap auf. Der E-Container "wusste" natürlich nichts von diesem Zusammenhang, er arbeitet ja nur mit dem SENTRAX Kookkurrenzen-Netz, das die Engine aus allen E-Dokumenten berechnet hat.

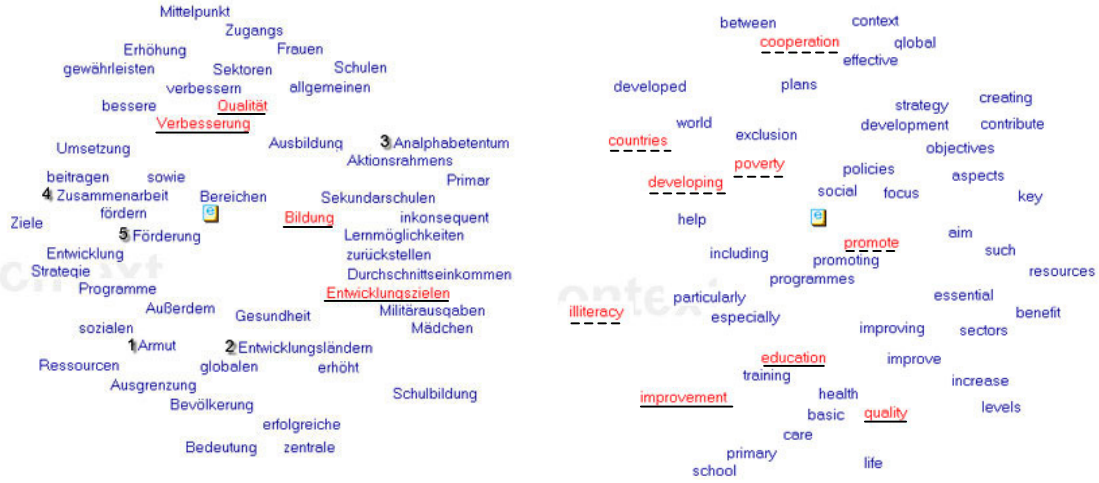


Abbildung 13 Links: Die D-ContextMap, die Unterstreichungen zeigen die Terme der Anfrage. Die nummerierten Wörter werden zur Übersetzung ausgewählt. Rechts: Die E-ContexMap, die gestrichelten Unterlinien bezeichnen die zufällig ausgewählten Übersetzungen.

Da in der ContextMap unterschiedliche Konzeptgruppen vorkommen, kann man sein eigenes Konzept aus verschiedenen Gruppen neu zusammenstellen und die verfolgte Idee interaktiv verfeinern, wie im obigen Beispiel. Die zusätzlichen Begriffe verstärken das jeweilige Konzept. In diesem Beispiel werden die Attribute „Armut - poverty“,

„Entwicklungsländern – developing countries“, „Analphabetentum - illiteracy“, „Zusammenarbeit - cooperation“ und „Förderung - promote“ zur Übersetzung herangezogen, weil diese Begriffe wegen ihres vermutlichen Zusammenhanges mit dem obigen Thema geeignet sind. Wegen der Annahme, dass es sich in dem Ausgangsdokument um das Thema „Qualitätsverbesserung und Förderung der Bildung in Entwicklungsländern“ dreht, wird erwartet, dass die übersetzten Attribute zu demselben Thema bzw. zu dem parallelen Dokumentpaar führen.

Anhand der deutschen Anfrage trifft die Suche das Dokument „De01/ep-01-09-06.txt13“, während die Addition der ins Englische übersetzten Attribute die SENTRAX zum Dokument „En01/ep-01-09-06.txt12“ führt. Dieses Dokument ist der parallele Text zu dem deutschen Dokument „De01/ep-01-09-06.txt13“.

Nachfolgend steht ein Auszug aus dem einzigen 100%-Trefferdokument in D mit der Anfrage *Entwicklungsziele, Qualität, Verbesserung, Bildung* :

Zugang von Kindern zur Bildung in den Entwicklungsländern

Der Präsident:

Nach der Tagesordnung folgt der Bericht (A5-0278 / 2001) von Frau Kinnock im Namen des Ausschusses für Entwicklung und Zusammenarbeit über die Grundbildung in den Entwicklungsländern im Kontext der Sondertagung der Vollversammlung der Vereinten Nationen über Kinder im September 2001 (2001 / 2030 (INI))

Kinnock:

(EN) Herr Präsident, ... Lassen sie mich die Ziele aufzeigen : unentgeltliche und obligatorische Bildung für alle , Halbierung des Analphabetentums bei Erwachsenen bis 2015 , Beseitigung geschlechtsbezogener Ungleichheiten an Primar- und Sekundarschulen bis 2005 und Ausweitung der Lernmöglichkeiten für Erwachsene und Jugendliche sowie Verbesserung der tatsächlichen Qualität der gebotenen Bildung . Wir sind dessen recht überdrüssig, von Entwicklungszielen zu hören. Sie werden gesetzt, und sie sind endlos. ...

Das gefundene (und hier tatsächlich parallele) Dokument aus E mit der Anfrage *quality, improvement, education, poverty, developing countries, illiteracy, cooperation, promote* ist:

Basic education in developing countries

President:

The next item is the report (A5 0278 / 2001) by Mrs Kinnock , on behalf of the Committee on Development and Cooperation, on basic education in developing countries in the context of the United Nations General Assembly Special Session on Children in September 2001 [2001 / 2030 (INI)] .

Kinnock:

Mr President, … Let me set out what the targets are : free and compulsory education for all , halving adult illiteracy by 2015 , eliminating gender disparities in primary and secondary schools by 2005 and extending learning opportunities for adults and young people and improvements in the quality of education provided . We are quite tired of hearing about development targets. They are set and they are endless. …

Fazit

Die Übersetzung der Anfragebegriffe ist oft nicht ausreichend für die bilinguale Suche, weil sie das Konzept nicht trägt, nicht bewahrt. Wegen der Mehrdeutigkeit wird das Konzept bei der Übersetzung häufig abgelenkt. Dadurch bleibt eine Suche oft erfolglos. Mit Hilfe der grafischen Darstellung von Attributen kann man gut überblicken, welches Thema die Attribute beschreiben. Wenn man das Thema des Dokumentes im Voraus wüsste, könnte man einen besseren Zusatzbegriff zu den beschreibenden Attributen finden. Auch bei den übersetzten Attributen können die Wörter als die Attribute des Begriffs in der Zielsprache ausgewählt werden. Bei der automatischen übersetzten Attributauswahl könnte die Wortbeziehung und die graphische Darstellung behilflich sein, um die nicht zu dem gesuchten Begriff gehörenden Attribute wegzuräumen. Diese Reinigung durch die Kookkurrenzen kann das Problem der Mehrdeutigkeit bei der Übersetzung aushebeln.

An diesem Beispiel kann man erkennen, dass die Übersetzung nicht immer für ein deutsches Wort funktioniert. Insbesonders erzeugt das deutsche Kompositum dieses Problem. Das deutsche Kompositum, z.B. „*Entwicklungsziel*“, entsteht aus zwei oder mehr Wörtern, die kombiniert wurden. Meistens findet man dieses nicht im normalen Wörterbuch. In dieser Situation zeigt sich der Vorteil des Konzeptnetzes, da die anderen übersetzbaren Wörter aus der Umgebung das Konzept bewahren können. Obwohl nicht

alle Attribute aus der Ausgangsprache übersetzt werden können, bieten die umgebenden Attribute in der Ausgangsprache eine Möglichkeit zur Ergänzung. Durch eine geschickte Auswahl der zusätzlichen Attribute zwecks Übersetzung kann das parallele Dokument gefunden werden, ohne das unübersetzbare Kompositum übersetzen zu müssen.

5.1.2 Hilfen aus der Umgebung

Die umgebungsbezogenen Begriffe können das Suchkonzept verschärfen. Zwei Methoden zur Auswahl, um die Zusatzbegriffe aus der Umgebung zu bekommen, sind die manuelle Auswahl vom Nutzer und die automatische Auswahl von den durch die ContextMap erzeugten bezogenen Begriffen. Bei der manuellen Auswahl wird das Geschehen der Zwischenstufen beobachtet. Die automatische Auswahl kann die Annahme der ersten *n* engen Begriffe oder die vertretenden Begriffe aller Gruppen sein. Weil die Auswahl von den vertretenden Begriffen aller Gruppen die technische Unterstützung durch den Programmcode verlangt, was zunächst noch vermieden werden soll, wird nur die Annahme der ersten *n* engen Begriffe getestet. Die Methoden müssen daraufhin verglichen werden, welche den Suchprozess am besten unterstützen kann.

Experiment

Ein weiteres Ziel ist, die beiden SENTRAX-Container in den unterschiedlichen Sprachen im Suchprozess zu vergleichen, um zu sehen, wie das Verhalten in den Zwischenstufen ist. Die meisten Nutzer verwenden nur zwei oder drei Suchwörter als Anfrage. Im folgenden Beispiel werden als D-Anfrage zunächst nur zwei Terme benutzt, und zwar *Lebensmittel* und *Gentechnik*. Die englische Übersetzung der deutschen Anfrage, nämlich *food* und *genetic engineering* wird in die E-SENTRAX eingegeben. Die beiden ContextMaps D und E bilden jeweils eine Begriffswolke. Auf dem Bildschirm sieht man, welche Wörter bzw. Begriffe einen Zusammenhang mit der Anfrage haben und welches Thema dazu gehört. Wenn man das Wort *Vorteile* noch zur D-Anfrage hinzufügt, führt dies hier zum Thema „Vorteile der Gentechnik für Lebensmittel“. Das Wort „*Nachteile“* ist dabei sehr nah. Das zugehörige Thema ist möglicherweise die Diskussion über die Vor- und Nachteile der Gentechnik. Mit den drei Suchwörtern erhalten wir

zwei passende Dokumente, ep-01-03-15.txt3 und ep-01-03-15.txt4, während bei der E-Suche drei Dokumente ermittelt werden, ep-01-03-15.txt3, ep-01-09-05.txt6 und ep-01-03-15.txt4. Ein zusätzlicher Begriff, *biotechnology,* wird bei der E-ContextMap hinzugefügt, weil in der D-ContextMap das Wort *Biotechnologie* ebenfalls vorkommt. Nach dieser Erweiterung durch diesen Begriff aus der Umgebung werden die beiden parallelen Dokumente getroffen.

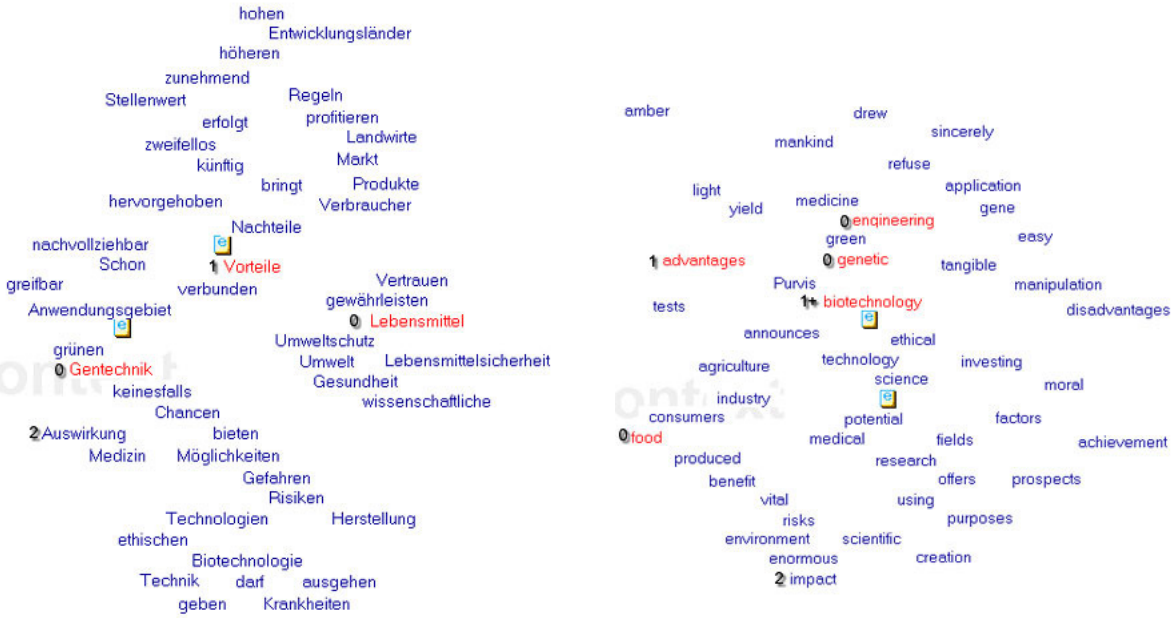


Abbildung 14 Links: die deutsche ContextMap mit der Anfrage „Gentechnik, Lebensmittel“. Rechts: die englische ContextMap mit der Anfrage „genetic, engineering, food“. Die Nummer „0“ bezeichnet den anfänglichen Zustand, „1“ und „2“ den folgenden Zustand und „1+“ bezeichnet die englische zusätzliche Ergänzung.

Wenn man sein Konzept verfeinern möchte, steht, z.B. für „Konsequenz der Gentechnik“ oder für „Nebenwirkung“, das Wort „*Auswirkung*“ in der ContextMap bereits zur Verfügung. Auch bei der englischen ist das Wort „*impact*“, das eine Übersetzung von *Auswirkung* ist, bereits zu sehen. Man kann erkennen, dass man den Zusammenhang zwischen seiner Vorstellung, den Begriffen und dem umgebungsbedingten Thema mit der SENTRAX ContextMap-Funktion bilden und daraus lernen kann, das eigene Konzept besser zu beschreiben. Außerdem eignen sich die Begriffswolken in den beiden Sprachen zur Realisierung der Suchabsicht. Anschließend ein Ausschnitt des D-

Dokuments ep-01-03-15.txt3 entsprechend den Suchwörtern „*Gentechnik*“, „*Lebensmittel*“, „*Vorteile*“, „*Auswirkung*“ :

....Dem Berichterstatter zufolge haben biotechnologische Entwicklungen positive Auswirkungen auf den Umweltschutz, die Qualität von Lebensmitteln, das Gesundheitswesen sowie für die Entwicklungsländer. Ob die Entwicklungsländer davon profitieren ist allerdings mehr als fraglich....

Wie im Anhang erläutert (siehe Abschnitt 7.2.1.2), lässt sich die ContextMap-Funktion auch in Form einer Liste ausgeben, das sähe bei unserem Beispiel dann so aus:

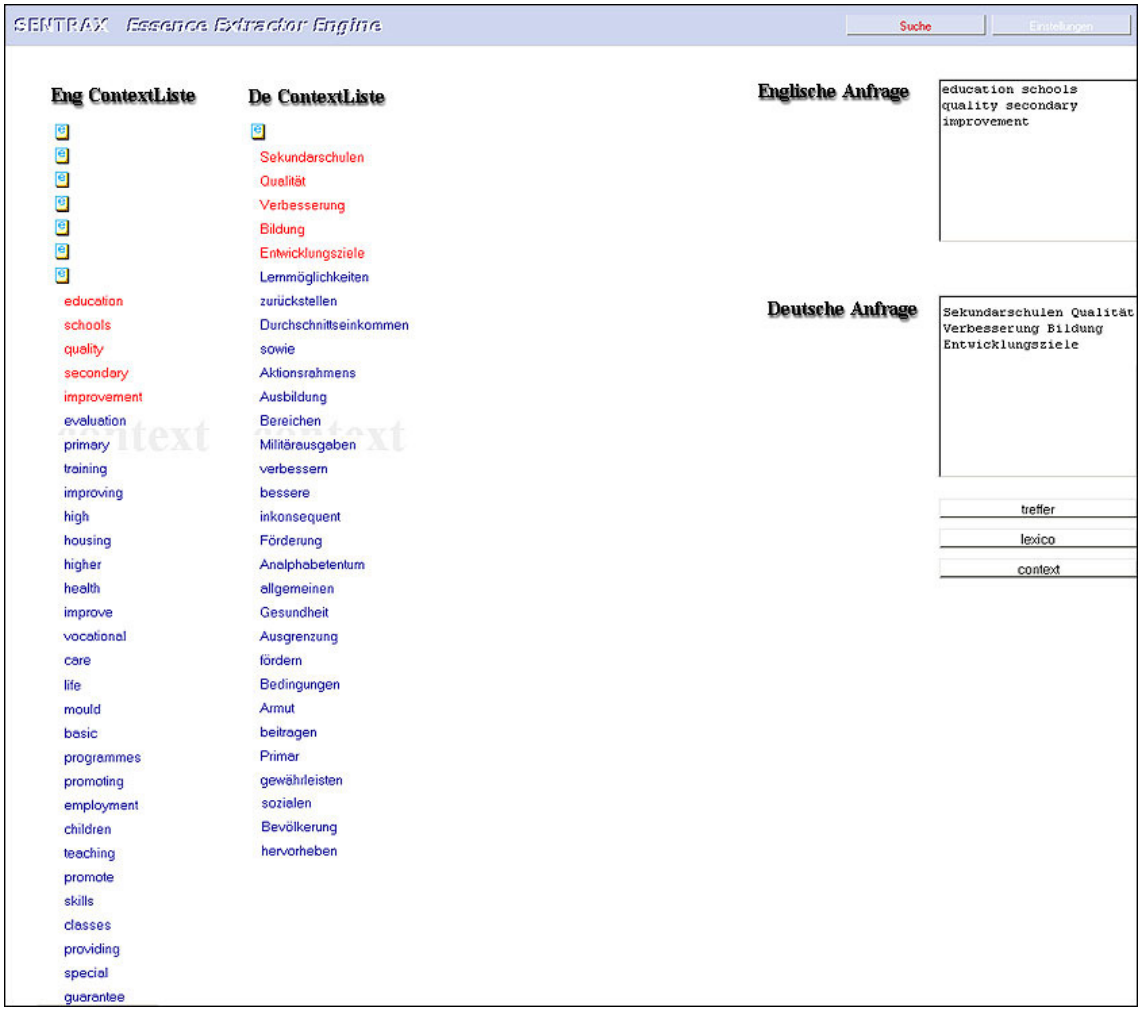


Abbildung 15 Die obersten r Begriffe nennen wir *r-Top Menge*. Diese Einträge sind bereits nach Kookkurrenzstärke sortiert. Die Grafik zeigt 30 Top Menge englische und deutsche ContextListe.

Die automatische Auswahl einer von der ContextMap erzeugten Wortliste wird hier untersucht. Die Untersuchungen teilen sich in zwei Abschnitte. Zunächst werden die r-Top Begriffe aus der Ausgangsprache übersetzt. Die Übersetzungen werden in dem Zielcon-

tainer mit den übersetzten Suchwörtern zusätzlich eingegeben. Die getroffenen Dokumente in der Zielsprache werden auf der Trefferliste bezüglich ihrer Rangfolge beobachtet, zunächst mit zwei zusätzlichen Übersetzungen, danach mit fünf und dann mit weiteren fünf Zusätzen. Obwohl das parallele Paar sehr früh durch die Übersetzung der Top-Begriffe erreicht wird, zeigt sich, dass man nicht garantieren kann, das parallele Paare durch die Verschärfung konvergieren.

Danach werden 30-Top bzw. 50-Top Begriffe der Zielsprache mit den 30-Top Begriffen der Ausgangsprache zusammen betrachtet. Wenn eine Übersetzung der 30-Top Begriffe der Ausgangsprache irgendeinen der 30-Top bzw. 50-Top Begriffe der Zielsprache trifft, wird der getroffene Begriff den Suchwörtern hinzugefügt. Alle getroffenen Zusatzbegriffe und die übersetzten Suchwörter werden auf dem Zielcontainer angewendet. Das Ergebnis eines Beispiels befindet sich in Tabelle 8.

		Anzahl der Zusatzbegriffe	TrefferDoc (englisch)
Anfang		0	ep-01-09-06-12 + 6 Dateien (100%)
Von 30-Top	1. Runde	5	4 Dateien (100%) ep-09-06-12 (63%)
	2. Runde	1	4 Dateien (100%) ep-09-06-12 (63%)
Von 50-Top	1. Runde	7	2 Dateien (100%) ep-09-06-12 (63%)
	2. Runde	6	Keine Datei (100%) 6 Dateien (63%) ep-09-06-12 (weg von ersten 30)

Tabelle 8 Ein Ergebnis der automatischen Auswahl entsprechend der anfänglichen Anfrage „Entwicklungszielen, Bildung, Qualität, Verbesserung, Sekundarschulen“ bzw. „education, quality, improvement, secondary schools“. Das parallele Zieldokument ist ep-01-09-06-12.

Fazit

Diese Untersuchung zeigt, dass die umgebungsbedingten Attribute das Suchkonzept erfüllen können, zumindest bei manueller Auswahl. Ohne Hilfe der miteinander verwobenen Attribute muss man den zusätzlichen Begriff selbst heraussuchen, um sein Konzept deutlich zu machen. Wahrscheinlich benötigt der Nutzer einige Zeit, um ein geeignetes Wort im Kopf zu finden. Insofern sind die umgebungsbedingten Attribute sehr hilfreich

um die Nutzeridee abzubilden. Außer bei der Verstärkung des Konzeptes hilft die Umgebung dem Nutzer auch bei der Prozessverfolgung, ob er in die richtige Richtung geht. Die Verfolgung wird durch die Beobachtung über das Konzeptnetz gelingen, je nach dem ob aktuelle umgebungsbedingte Attribute das Suchkonzept verstärken oder abschwächen. Durch die Prozessverfolgung mittels Beobachtung des Umgebungsverhaltens kann man sicherstellen, dass die Suche ihr Ziel nicht verfehlen wird, zumindest bei der manuellen Auswahl der Attribute.

Man kann auf der ContextMap erkennen, ob die Übersetzung der Attribute in der Zielsprache erfolgreich ist. Falls die Attribute in dem selben Konzept in der Ausgangsprache wie auch in der Zielsprache vorkommen und sie die ihnen entsprechenden Dokumente aufzeigen, dann kann man in der Zielsprache durch das Hinzufügen von weiteren, mit je einem Attribut in der Ausgangsprache vergleichbaren, Attributen die Trefferliste soweit verschärfen, bis nur noch das parallele Dokument auf dieser erscheint. Dies kann aber auch bei der automatischen Auswahl der zugehörigen Attribute funktionieren.

Ein Hauptproblem der automatischen Auswahl von Begriffen ist die Divergenz. Es gibt bisher aber keine Anzeige eines Hinweises, welche Begriffe auf dem Konzeptnetz der Zielsprache zur richtigen Antwort führen. Wenn man Zusatzbegriffe willkürlich (stumpf automatisch) hinzubringt kann die gewünschte Suchabsicht verwischt werden. Wenn falsche Begriffe in der Anfrage starken Einfluss haben, werden die bilingualen Paare verfehlt. Grund dafür ist, dass die Zusatzbegriffe mit der selben Priorität wie die Suchbegriffe betrachtet werden. Sie haben Zusammenhänge in dem Konzeptnetz, aber vielleicht nicht in dem selben Dokument. Außerdem ergibt sich aus dem Unterschied in der sprachlichen Nutzung bzw. im Schreibstil möglicherweise eine erfolglose automatische Auswahl. Wenn der Übersetzer bzw. der Verfasser mit einem sehr individuellen Stil schreibt, wirkt sich dies sowohl auf den Zusammenhang der Begriffe, als auch auf die automatische Übersetzung durch das elektronische Wörterbuch aus. Dies geschieht, weil die Gestalt der bilingualen Konzeptnetze durch den persönlichen Stil und die sprachliche Nutzung, z.B. auf Deutsch mit dem Nomen aber auf Englisch mit dem Verb oder der Mehrwortgruppe, sich unterschiedlich darstellt und die allgemeine Übersetzung durch das Wörterbuch die individuellen Wörter nicht ausgleichen kann. Es bleibt

zu prüfen, ob kommende Versionen der bilingualen SENTRAX, die mit umfangreicheren Vorverarbeitungsfunktionen ausgestattet werden, hier bessere Ergebnisse liefern.

Außerdem wirkt sich der zu den Suchwörtern zusätzlich addierte 1-Top Begriff so aus, dass sich die restlichen Begriffe in der ersten n-Top Wortliste meist nur umgruppieren. Manchmal ergeben sich aber auch neue Begriffe in der neuen n-Wortliste. Diese positionieren sich dann aber meist am Listenende.

5.1.3 Ähnliche Dokumente

Hier wird überprüft, ob die SimilarDoc-Funktion der SENTRAX die parallelen ähnlichen Dokumente finden kann.

Experiment

Die SimilarDoc-Funktion sucht die zu einem ausgewählten Trefferdokument ähnlichen Dokumente in der Datenbasis. Zusätzlich liefert sie eine prozentuale Übereinstimmung zum gewählten Referenzdokument. Für uns interessant ist, ob man in D ähnliche Nachbarn erhält wie in E. Die Tabelle unten zeigt, dass die SimilarDoc-Funktion tatsächlich einander parallele Dokumente herausarbeitet. In der ersten Zeile sind die beiden parallelen Dokumente notiert, die jeweils unabhängig voneinander in ihren Containern mit der SimilarDoc-Funktion aktiviert wurden.

	D-Index	E-Index	D-Rangplatz zum E-Dokument
0	**01-03-15-3**	**01-03-15-3**	jeweils aktivierte Dokumente
1	01-02-13-11	01-02-13-11	1
2	01-02-13-8	01-02-13-8	2
3	01-11-14-2	01-03-15-4	4
4	01-03-15-4	01-11-29-2	5
5	01-11-29-2	01-11-14-2	3
6	01-05-30-4	01-02-15-1	7
7	01-02-15-1	01-10-04-10	9
8	01-02-14-4	01-05-30-4	6
9	01-10-04-10	01-03-14-3	11
10	01-06-12-4	01-11-29-3	15

Tabelle 9 Dokumente und Rangplätze gemäß SimilarDoc-Funktion

Die erste Spalte zeigt die Reihenfolge der zum ersten ähnlichen Dokumente in D. Die zweite Spalte analog dazu für E. Die dritte Spalte zeigt den Rangplatz des D-Dokuments, das dieselbe Nummer wie das E-Dokument hat, zu diesem also parallel ist. Man kann erkennen, dass die Umgebung in Bezug auf "gleichartigen Inhalt" durch die SimilarDoc-Funktion weitestgehend erhalten bleibt.

Fast alle ähnlichen Dokumente auf den ersten zehn Rangplätzen handeln vom Thema Genforschung. Die Dokumente auf den ersten zwei Plätzen sind „die Freisetzung genetisch veränderter Organismen“. Andere Themen sind Humangenetik, Forschung und technologische Entwicklung. Der folgende Ausschnitt ist aus der deutschen Datei ep-01-02-13.txt11 mit dem Thema „Freisetzung genetisch veränderter Organismen (Fortsetzung)“

Boudjenah:

Herr Präsident, das Vorhandensein von GVO in unseren Nahrungsmitteln und die Ungewissheiten hinsichtlich deren Auswirkungen auf den Menschen und seine Umwelt sind nunmehr öffentliche Fragen, und das ist gut so. ...

Diamantopoulou:

... Bis Juni 2001 wird die Kommission Vorschläge zur Kennzeichnung vorlegen, die dem Verbraucher mehr Informationen über Lebensmittel aus GVO vermitteln. Insbesondere beabsichtigen wir, den gegenwärtigen Ansatz fallen zu lassen, wonach das DNS-Protein das entscheidende Kriterium ist. Auf diese Weise haben die Verbraucher eine maximale Auswahl zwischen genetisch veränderten und konventionellen Erzeugnissen. Unserem Vorschlag zufolge werden auch verarbeitete Lebensmittel in das Kennzeichnungssystem einbezogen sein. ...

Fazit

Wie erwartet liefert die SimilarDoc-Funktion fast identische parallele Paare. Der Unterschied liegt nur im Rangplatz. Dies ist für den Nutzer sehr hilfreich, der ein richtiges Dokumentpaar finden und andere entsprechende Dokumente ermitteln will.

Weil die aktuelle SimilarDoc-Funktion mit der Wortmusterabgleichung arbeitet, kann sie bei der normalen Situation der krosslingualen Suche schwach sein und eventuell keine semantische Ähnlichkeit liefern. Die Ähnlichkeit der Dokumente orientiert sich ja an der Zahl der gleichen Wörter und nicht an ihrer Bedeutung. Diese Methode funktioniert für übliche parallele Korpora, aber für vergleichbare Korpora eventuell nicht,

weil die Ähnlichkeit nur auf Wortebene liegt. Eine Erweiterung auf der Semantikebene ist erforderlich, damit die autorspezifischen Schreibweisen sowie die Homonymen und Synonyme erkannt werden können.

5.1.4 E→D Suche

Der Suchbedarf beschränkt sich nicht nur auf eine Richtung. Hier wird gezeigt, dass die bilinguale Suche mittels Konzeptnetz auch in der Gegenrichtung funktioniert.

Experiment

Die englische Anfrage in diesem Beispiel ist aus den Wörtern „energy“, „saving“, „ecology“, „environment“ und „research“ zusammengesetzt. Dieses Konzept führt zu dem E-Dokument „ep-01-06-13.txt11“. Die Übersetzung der Anfrage mit Hilfe des Online-Wörterbuchs ist „Energie“, „sparend“, „Ökologie“, „Umwelt“ und „Forschung“. Obwohl das Wort „sparend“ nicht gefunden werden kann, taucht das Wort „einzusparen“ in der deutschen Umgebung auf. Nach der Auswahl des zusätzlichen Attributs „einzusparen“ wird das deutsche parallele Dokument „ep-01-06-13.txt11“ getroffen.

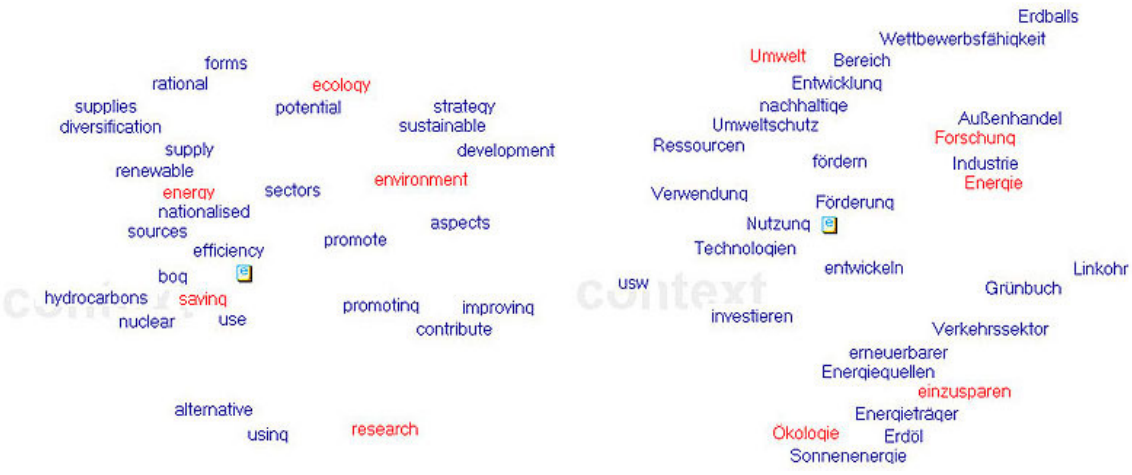


Abbildung 16 Links: die ContextMap in der Ausgangsprache Englisch. Rechts: die ContextMap in der Zielsprache Deutsch.

Fazit

Anhand dieses Beispiels kann man erkennen, dass die Suche in der Gegenrichtung (E→D) ebenfalls funktioniert. Bemerkenswert dabei ist, dass der Zusammenhang von den übersetzbaren Attributen andere Attribute hervorbringt, wie in diesem Beispiel das Wort „einzusparen“. Die anderen Übersetzungspaare sind natürlich enthalten, z.B. „promote – fördern“, „using (use) – Nutzung (Verwendung)“, „renewable – erneuerbarer“ usw. Das Wort „sectors“ kann vielleicht dem Wort „Bereich“ oder dem Wort „Verkehrssektor“ entsprechen, weil es auf Englisch allein stehen kann oder mit anderem Wort zusammen stehen kann.

5.2 Sonderfälle

Hier werden vier Fälle betrachtet: (1) der Zielcontainer ist viel größer als der Ausgangscontainer (2) der Zielcontainer ist kleiner als der Ausgangscontainer (3) das relevante Dokument wird im Zielcontainer entfernt (4) der Zielcontainer wird mit anderen, fremden Texten erweitert.

5.2.1 Großer Zielcontainer

Der Zielcontainer wird in dieser Situation durch weitere Dokumente ohne parallele Entsprechungen in dem Ausgangscontainer erweitert. Das heißt, alle Dokumente im Ausgangscontainer haben parallele Partner im Zielcontainer, aber nicht umgekehrt. Das Suchverhalten und das Konzeptnetz in der Zielsprache werden hier daraufhin beobachtet, ob sie wie im Standardfall gebildet werden können.

Experiment

Der E-Container beinhaltet eine Textsammlung aus den Jahren 2000 und 2001, während der Ausgangscontainer (D-Container) nur die Textsammlung aus dem Jahr 2001 besitzt. Es fängt an mit der deutschen Anfrage „Gentechnik, Lebensmittel, Vorteile“. Die deut-

schen Dokumente „ep-01-03-15.txt3“ und „ep-01-03-15.txt4“ werden mit 100 Prozent getroffen. Die englische Anfrage (wie in Abschnitt 5.1.2) „genetic engineering, food, advantages, biotechnology“ wird auf den E-Container angesetzt. Die 100-prozentigen Trefferdokumente sind nicht nur aus dem Jahr 2001 („ep-01-03-15.txt3“, „ep-00-10-25.txt2“), sondern auch aus dem Jahr 2000 („ep-01-03-15.txt4“). Dank der Kookkurrenz kann man ein zusätzliches Attribut aus der Wortumgebung in der Ausgangsprache heranziehen, um es in die Zielsprache zu übersetzen.

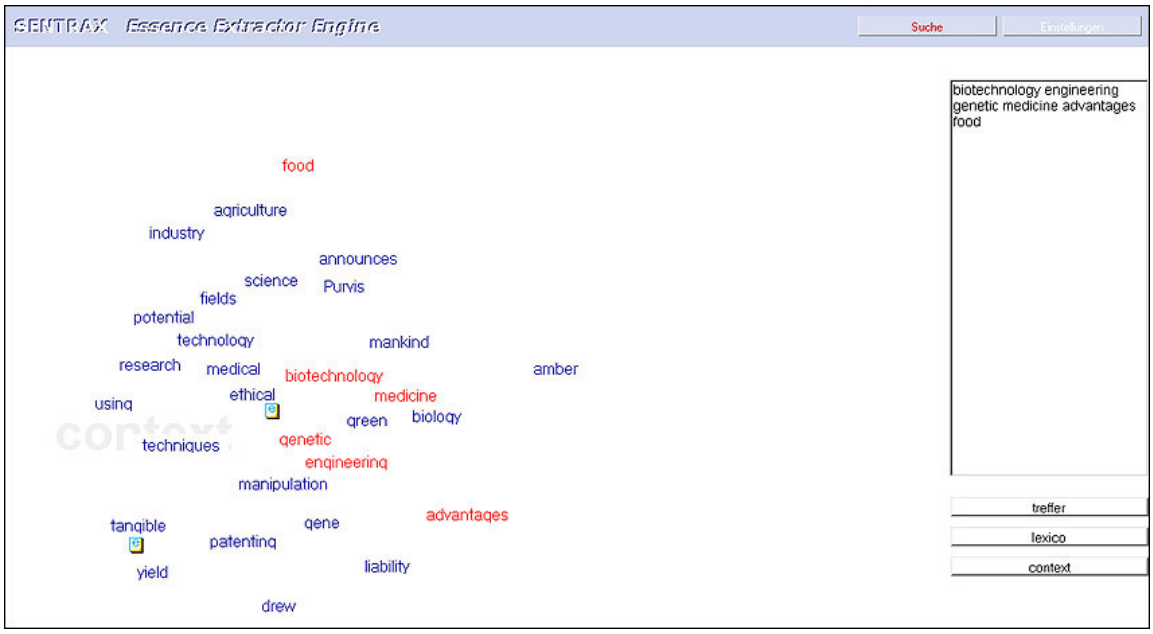


Abbildung 17 Die endgültige englische ContextMap auf dem E-Container aus den Jahren 2000 und 2001 mit dem Konzept der Anfrage „genetic engineering, food, advantages biotechnology, medicine“.

In diesem Beispiel wird das Wort „Medizin“ transferiert. Mit der zusätzlichen Übersetzung „medicine“ ergeben sich nur die Dokumente „ep-01-03-15.txt3“ und „ep-01-03-15.txt4“.

In einem zweiten Beispiel fängt es auf dem deutschen Container mit der Anfrage „USA, Ölindustrie, Klimawandel“ an. Mit nur zwei Attributen wird ein einziges 100-prozentiges Dokument ermittelt. Glücklicherweise sind die Übersetzungen der beiden Komposita („Ölindustrie ≡ oil industry“ und „Klimawandel ≡ climate change“) im Wörterbuch enthalten. Die übersetzte Anfrage ergibt 19 100%-Dokumente auf der Trefferliste. Im Gegensatz zum letzten Beispiel werden hier viele zusätzliche Attribute benö-

tigt, um ein einzelnes paralleles Dokument zu erlangen. Die Attribute werden von der Wortumgebung in der Ausgangsprache so gewählt, dass nur die Wörter, deren Übersetzung auf der ContextMap in der Zielsprache stehen, betrachtet werden. Die Zusatzterme sind „American, Protocol, Bonn, Kyoto, gas emissions effect, global“[25]. Diese Zusatzterme werden zusammen mit der anfänglichen übersetzten Anfrage in die Zielsprache eingegeben. Das parallele Dokument „ep-01-04-05.txt9“ bleibt übrig als einziger 100%-Treffer.

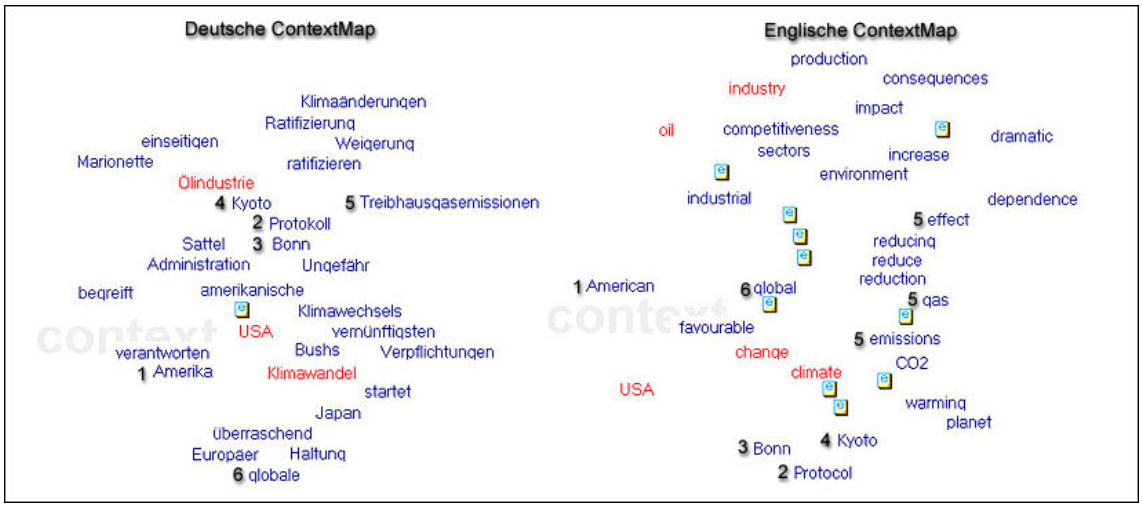


Abbildung 18 Vergleich der ContextMap zu der deutschen Anfrage „USA Ölindustrie Klimawandel“ und der englischen anfänglichen Anfrage „USA oil industry climate change“.

[25] gas emissions effect ist die Übersetzung des Wortes „Treibhausgasemissionen“ von http://dict.leo.org/

Fazit

Obwohl beide Beispiele einen doppelt so großen Zielcontainer als den Ausgangscontainer haben, erlauben sie eine erfolgreiche Suche. Sie unterscheiden sich allerdings in der Anzahl der nötigen zusätzlichen Attribute. Im ersten Beispiel wird nur ein Attribut benötigt, um das parallele Dokument herauszufiltern, während im zweiten Beispiel bis zu sechs Zusatzattribute nötig sind. Dabei ist anzumerken, dass sich in beiden Fällen die Übersetzungspaare bereits auf den beiden ContextMaps befinden.

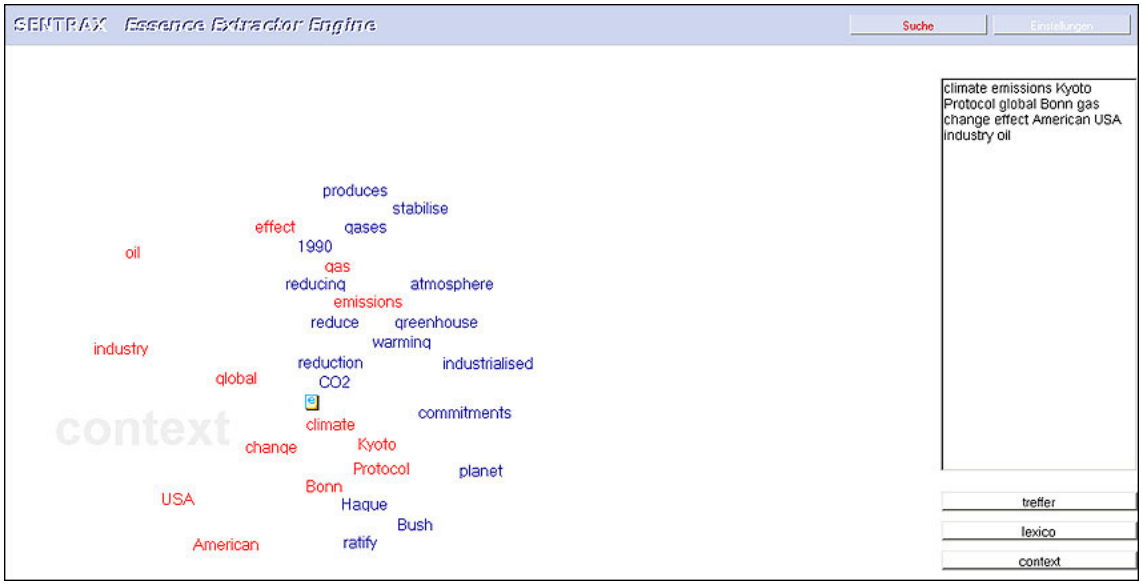


Abbildung 19 Die englische ContextMap mit den Zusatzattributen.

In dem getroffenen englischen Dokument aus dem Jahr 2000 „ep-00-10-25.txt2“ geht es um „food safety“. Es scheint, dass noch weitere ähnliche Dokumente im Zielcontainer gefunden werden können. Im Folgendem wird ein Ausschnitt aus dem E-Dokuments „ep-00-10-25.txt2“ entsprechend den Suchwörtern „*genentic engineering, food, advantages, biotechnology*“ gezeigt, an dem man erkennen kann, dass das Thema „food safety“ sehr nah an dem Konzept der Anfrage liegt und das Dokument somit als Treffer in Frage käme.

…the sheer number of food scandals and the debate about using genetic engineering in food production have undermined the present food safety system . Optimum food safety standards and falling consumer protection standards cannot be reconciled. …

… Another issue which will confront us over the coming months is that of GMOs in food production and we must be open to the potential of biotechnology. In this respect it could be a serious mistake to assume that biotechnology means poor quality or unsafe food. For example, GMO foods offer the opportunity to reduce the levels of pesticide residues and improve nutritional quality, it would be negligent to ignore these advantages. However, I fully support the introduction of clear, non-technical and standardised labelling in the context of GMO food products. Moreover, no food products which are genetically modified or contain genetically modified ...

Es kann sein, dass eine erfolgreiche Suche in der Zielsprache viel mehr Attribute benötigt als in der Ausgangsprache, weil das erforderliche Dokument in der Zielsprache durch weitere Attribute eingegrenzt werden muss. Begründet liegt dies in der Sprache selbst. Hier müsste man an weitere sprachspezifische Vorverarbeitungen denken.

Durch die umgebungsbedingten Übersetzungspaare kann die automatische Abgleichung gemäß des Ranges erfolgen, indem die Übersetzungen aller Attribute in der Ausgangsprache mit den Attributen in der Zielsprache abgeglichen werden. Wenn sie einander entsprechen, werden sie als Zusatzterme für die Anfrage in der Zielsprache hinzugefügt. Falls das Ergebnis nicht ausreichend gut ist, kann dieses Verfahren nochmals wiederholt werden.

5.2.2 Kleiner Zielcontainer

Hier wird im Gegensatz zu 5.2.1 der Fall „kleinerer Zielcontainer“ untersucht. Möglicherweise ergibt sich wegen der Dokumente ohne Partner ein anderes Konzeptnetz im Zielcontainer als im Ausgangscontainer.

Experiment

Der E-Container beinhaltet nur die Hälfte der Textsammlung aus dem Jahr 2001, während der Ausgangscontainer die Textsammlung aus dem ganzen Jahr 2001 enthält. Zwei Unterfälle sind zu unterscheiden: (1) der Abzug der Hälfte, in der die relevanten Dokumente nicht enthalten sind, (2) der Abzug der Hälfte, in der die relevanten Dokumente liegen.

Die Anfrage „genetic engineering, food“ wird auf dem E-Container in beiden Unterfällen anfänglich benutzt. Im ersten Fall werden die Dokumente aus dem zweiten Halbjahr abgezogen, wo kein 100-prozentiges relevantes Dokument enthalten ist. In diesem Zielcontainer sind nur 381 von 728 Dokumenten, wobei der Ausgangcontainer 731 Dokumente enthält. Die Attribute werden während der Suche daraufhin beobachtet, ob sie sich von denen aus dem ganzjährigen E-Container unterscheiden und ob sie zu denselben Dokumenten wie im ganzjährigen E-Container führen. Im zweiten Fall, wo der Zielcontainer 347 Dokumente aus dem zweiten Halbjahr beinhaltet, in dem kein 100-prozentiges Dokument entsprechend der obigen Anfrage vorkommt, wird analog verfahren.

Die vorkommenden Begriffe aus dem ersten Fall sind ziemlich ähnlich wie die im ganzjährigen E-Container (12 von 17 bezogene Begriffe), während sich nur 4 von 17 bezogenen Begriffen beim zweiten Fall im ganzjährigen E-Container befinden. Grund dafür ist wahrscheinlich, dass die Kookkurrenz zwischen der Anfrage und anderen Begriffen im ersten Halbjahr sehr stark ist, weil zwei der drei 100-prozentigen entsprechenden Dokumente aus dem ersten Halbjahr sind. Mit dem Zusatzattribut „advantages“ werden die 100-prozentigen Dokumente „ep-01-03-15.txt3“ und „ep-01-03-15.txt4“ aus dem ersten Halbjahr getroffen und das 100-prozentige Dokument „ep-01-09-05.txt6“ aus dem zweiten Halbjahr getroffen. Würde noch das Attribut „biotechnology“ wie in Abschnitt 5.2.1 hinzugefügt, würden nur die beiden Dokumente aus dem ersten Halbjahr auftauchen.

Abbildung 20 Die Kookkurrenzliste des ganzenjährigen E-Containers im Vergleich zu denen aus dem ersten und zweiten Halbjahr.

Fazit

Obwohl der halbjährige Zielcontainer halb so groß ist wie der Ausgangscontainer, funktioniert der Suchmechanismus dennoch und liefert noch dasselbe Ergebnis. Am Charakter der Attributumgebung kann man grob erkennen, um welches Thema es sich im Container drehen kann. Wenn man beispielweise die Attributliste beobachtet, kann man ungefähr erahnen, dass es sich im ersten halbjährigen Container um die Anwendung der Gentechnik und im zweiten halbjährigen Container um die wissenschaftliche Forschung dreht.

Bei der selben Anfrage entstehen auf zwei Wortumgebungen, die eine nicht leere Schnittmenge besitzen aber nicht identisch sind, aufgrund anderer Kookkurrenzverhältnisse innerhalb der einzelnen Umgebungen unterschiedliche Attributlisten. Weil die Kookkurrenzen nur in einem Container ermittelt werden, können sie den Inhaltscharakter des Containers repräsentieren. Die Stärke der Kookkurrenz hängt davon ab, wie oft die Wörter miteinander in demselben Kontext vorkommen. Im Ergebnis der Beispielan-

frage dominieren die Kookkurrenzwörter aus dem ersten Halbjahr bei dem ganzjährigen Container sehr deutlich, da sie sehr stark mit den Suchwörtern kookkurrieren.

Zum Zugriff auf die gewünschten Dokumente benötigt man noch weitere Attribute. Das Suchergebnis durch das Konzeptnetz auf dem kleinen Container ist ähnlich dem Suchergebnis aus dem normalen bzw. dem ganzenjährigen Container. Nur die Dokumente aus der anderen Hälfte fehlen. Der teilweise Abzug stört das Suchverhalten kaum. Solange die zu treffenden Dokumente in dem Container enthalten bleiben, findet man sie durch die selben Suchwörter

5.2.3 Abzug des relevanten Dokumentes

Hier wird untersucht, was passiert, wenn der Zielcontainer keinen parallelen Partner erhält. Wegen des Abzugs der parallelen Partner wird der Zusammenhang der Begriffe im Konzeptnetz verändert.

Experiment

Der Ausgangscontainer wird erst aus der gesamten Dokumentensammlung aus dem Jahr 2001 gebildet, während das Dokument „ep-01-03-15.txt3“ in dem Zielcontainer aus dem gleichen Jahr abgezogen wird. Das abgezogene Dokument ist eins von zwei Dokumenten, die der Anfrage „Gentechnik, Lebensmittel, Vorteile“ bzw. „genetic engineering, food, advantages“ entsprechen. Auf dem Zielcontainer bzw. E-Container wird das Attribut „biotechnology“ in der Suchanfrage hinzugefügt. Nach dem Suchprozess werden die Trefferlisten verglichen.

Dieselbe Anfrage wird für alle Untersuchungen verwendet, ohne die Auswahl während des Prozesses zu verfolgen, weil der Abzug nur auf dem Zielcontainer erfolgt.

Weil es zwei entsprechende Dokumente („ep-01-03-15.txt3“ und „ep-01-03-15.txt4“) auf die Anfrage „genetic engineering, food, advantages, biotechnology“ gibt, werden zunächst ein und später zwei entsprechende Dokumente aus der Sammlung abgezogen. Die Trefferlisten beider Abzugfälle werden mit dem ursprünglichen E-Container vergli-

chen. Natürlich fallen die abgezogenen Dokumente aus den neuen Dokumentlisten heraus. Die Dokumente auf den weiteren Rangplätzen rücken in der Reihefolge im Vergleich zu der ursprünglichen Dokumentliste entsprechend um einen Platz nach vorne. Dieses ergibt sich analog beim Abzug der beiden entsprechenden Dokumente.

Die Wortlisten in der ContextMap werden mit der ursprünglichen Liste verglichen. Anzumerken ist, dass sich die umgebungsbezogenen Attribute gemäß derselben Anfrage wegen des Abzugs verändern. Die Attribute „research“, „purposes“ und „factors“, die nicht in der vollen Liste vorkommen, tauchen dagegen auf den Abzugslisten auf. Es gehen einige Attribute verloren, weil die Kookkurrenzhäufigkeiten zwischen den Attributen und den Suchwörtern durch die abgezogenen Dokumente geschwächt werden. Dadurch werden die Attribute der übrigen Dokumente auf der Wortliste ebenfalls beeinflusst. Die Veränderung der Kookkurrenzhäufigkeit verursacht eine Variation des Konzeptnetzes.

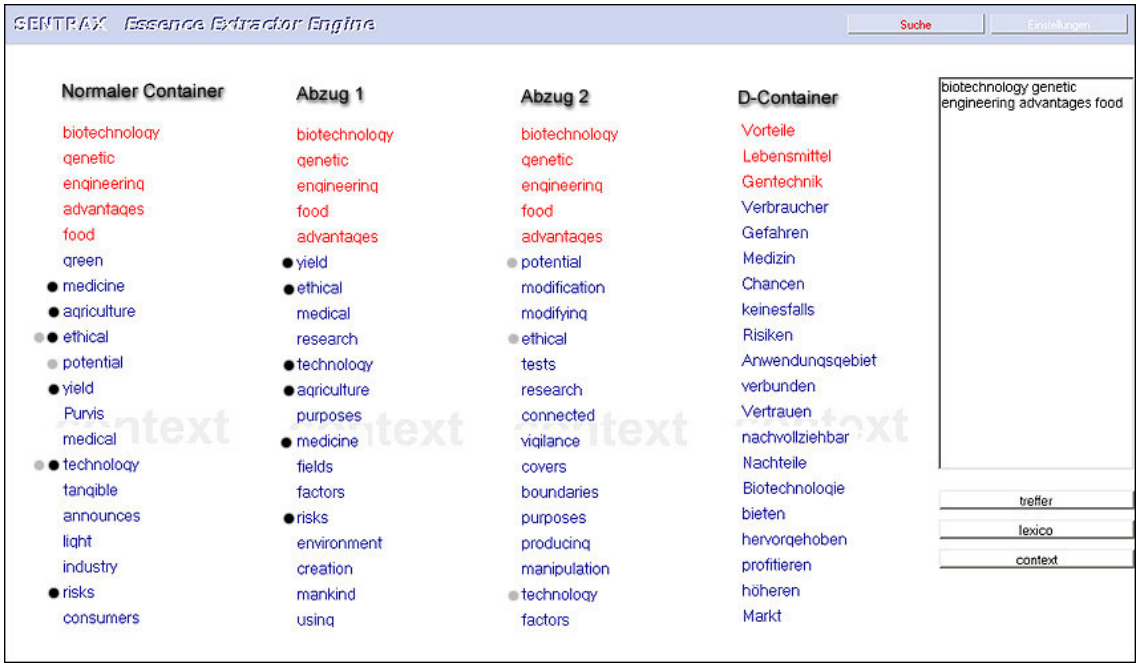


Abbildung 21 Die Wortliste der ContextMap: der normale Container repräsentiert die englische volle Sammlung, Abzug 1 repräsentiert die englische Sammlung mit einem abgezogenen entsprechenden Dokument bzw. Abzug 2 die englische Sammlung mit zwei abgezogenen entsprechenden Dokumenten, D-Container repräsentiert die deutsche volle Sammlung.

Der größere Container wird gebildet, indem die Dokumentensammlungen aus zwei bzw. drei Jahren in einem Container zusammengelegt werden. Der Abzug der Dokumente wird wie in der Untersuchung eines Jahres durchgeführt. Die bezogenen Begriffe werden daraufhin betrachtet, wie sie sich bei einem großen Container und kleinen Container ergeben bzw. verschieben. Im Folgenden werden Prozentzahlpaare dargestellt, die wie folgt gebildet wurden:

i. Die Anzahl der gleichen Umgebungsbegriffe von komplettem Container und Abzug1.

ii. Die Anzahl der gleichen Umgebungsbegriffe von komplettem Container und Abzug2.

Das Paar hat auf dem Ein-Jahres-Container die Werte (40, 20), auf dem Zwei-Jahres-Container die Werte (60, 27) und auf dem Drei-Jahres-Container die Werte (67,53).

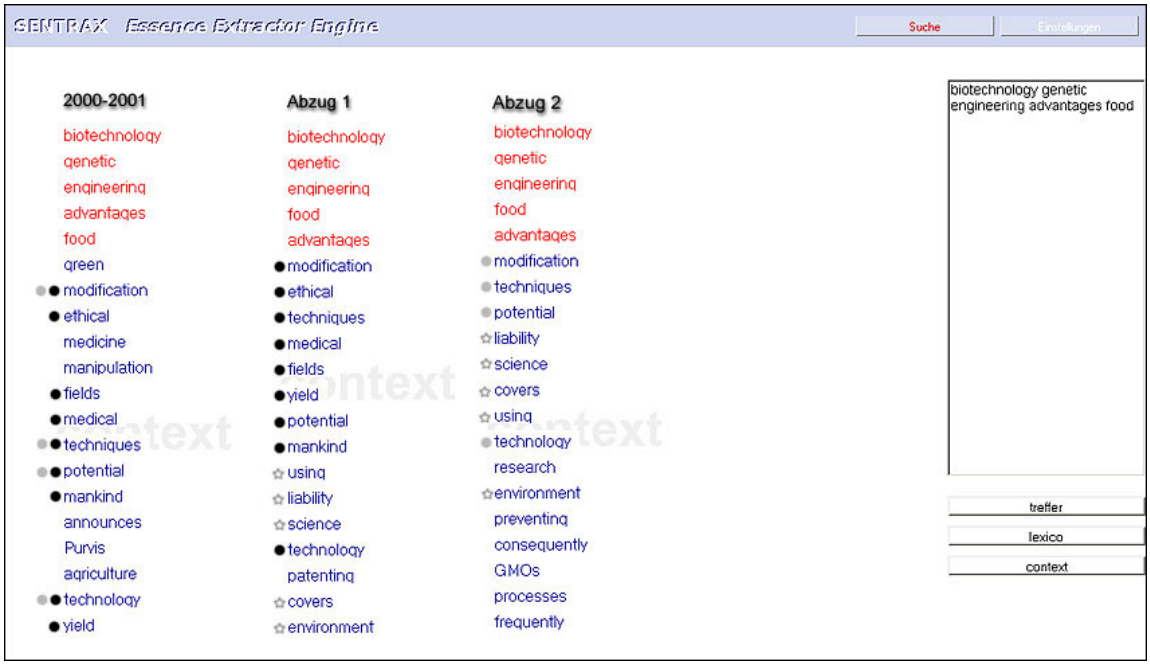


Abbildung 22 Die Wortliste der ContextMap-Funktion: der englische Container stammt aus dem Jahr 2000-2001. Der Abzug1 ist der englische Container abzüglich des Dokuments „ep-01-03-15.txt3“ und Abzug2 ist der englische Container abzüglich der Dokumente „ep-01-03-15.txt3“ und „ep-01-03-15.txt4“. Die Listen wurden durch die eingegebenen Suchwörter „biotechnology, genetic engineering, advantages, food“ erzeugt. Nur die ersten 20 Wörter auf der Rangliste werden hier gezeigt.

Aus der Abbildung 22 ist zu erkennen, dass (neben den roten) weitere fünf Begriffe beim Abzug1 und beim Abzug2 gleich sind (gekennzeichnet durch grauen, transparenten Stern). Diese fünf befinden sich nicht im kompletten Container. Diese Begriffe werden von unten nach oben hochgezogen, weil frühere obere Begriffe durch den Verlust an Stärke im Rang absteigen.

Bei dem Drei-Jahres-Container ist es ebenfalls so, dass einige Begriffe bei Abzug 1 und bei Abzug 2 auftauchen, aber es gibt nur vier gleiche Begriffe auf den ersten zwanzig Rangplätzen. Einige der vier Begriffe des Drei-Jahres-Containers liegen aber auch auf der Liste aus dem Zwei-Jahres-Container.

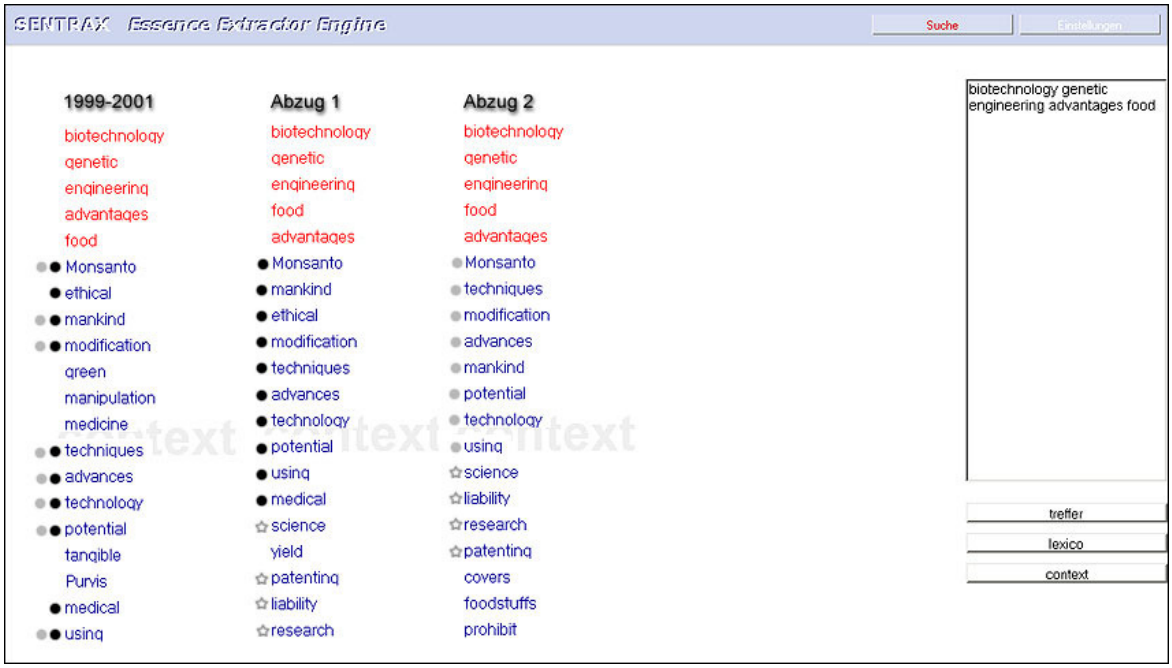


Abbildung 23 Die Wortliste der ContextMap: der englische Container stammt aus dem Jahr 1999-2001. Der Abzug 1 ist der englische Container abzüglich des Dokuments „ep-01-03-15.txt3“ und Abzug 2 ist der englische Container abzüglich der Dokumente „ep-01-03-15.txt3“ und „ep-01-03-15.txt4“. Die Listen wurden durch die eingegebenen Suchwörter „biotechnology, genetic engineering, advantages, food“ erzeugt. Nur die ersten 20 Wörter auf der Rangliste werden hier gezeigt.

Fazit

Obwohl einige relevante Dokumente abgezogen werden, beeinflusst dies die Reihenfolge der getroffenen Dokumente in der Liste bei der Suche mit derselben Anfrage nicht. In der Dokumentenliste ändern sich die Platzierungen, indem die Treffer auf den nächsten Rangplatz nach vorne geschoben werden, die nach den abgezogenen Dokumenten platziert waren. Aufgrund der aktivierten Suchwörter bleibt der Rest an entsprechenden Dokumenten weiterhin erhalten. Es wird klar deutlich, dass in dem Konzeptnetz bzw. der Wortliste die Beziehungen wegen des Abzugs verändert werden. Diese Veränderung deckt auf, wie die Orientierung des suchenden Themas durch die Suchwörter bewusst abgelenkt werden kann.

Eine deutliche Veränderung des Konzeptnetzaussehens durch den Abzug einiger Dokumente ergibt sich, wenn der Container klein ist. Die Veränderung des Konzeptnetzes bei gleicher Anfrage mit dem Abzug von einigen 100%-Dokumenten ist nicht so dramatisch. Dies gilt vor allem, wenn sehr viele relevante Dokumente im Container existieren.

5.2.4 Zwei Sprachen in einem Container

Die Hypothese ist hier, dass die Textsammlung zwei oder mehrere Sprachen in einem Container enthalten kann und das Konzeptnetz bzw. die abgerufenen Dokumente entsprechend der Sprache der Anfrage ermittelt werden können.

Experiment

Die deutsche und englische Textsammlung werden zusammen in den gleichen Container gepackt. Die üblichen deutschen und englischen Anfragen werden auf diesem Container getestet. Die Ergebnisse der Suche erscheinen ganz normal wie bei getrennten Containern. Die relevanten Dokumente werden je nach der Sprache der Anfrage ermittelt. Fraglich ist, ob die transliterierten Wörter ein Problem auf dem Konzeptnetz durch die SENTRAX erzeugen. Man findet beispielweise das Wort „Situation“ sowohl im deutschen als auch im englischen Text.

Hätten die bilingualen Texte im gleichen Container gelegen, wären die Assoziationen von einer Sprache zu der anderen Sprache aufgrund der transliterierten Wörter entstanden. Tauchen gemischte Begriffe aus mehreren Sprachen auf dem Konzeptnetz auf, wenn der Container aus zwei oder mehr Sprachen erzeugt wurde? Dies ist zu testen. Das Wort „USA“ wird als Anfrage eingegeben. Das Konzeptnetz wird durch die Funktion „ContextMap“ errechnet. Erstaunlicherweise tauchen viele deutsche Begriffe auf. Dagegen kommen nur wenige englische Begriffe vor. Es wird weiter mit dem deutschen „Klima“ und englischen „climate“ getestet.

Wird die Suchanfrage „USA, Klima“ im gemischten Container eingegeben, befindet sich der englische Begriff nicht mehr auf dem Konzeptnetz. Genau so beeinflusst das englische Zusatzwort in der Suchanfrage „USA, climate“ das Konzeptnetz für die englischen Begriffe. Wenn die Suchanfrage aus beiden Sprachen gebildet wird, „USA, climate, Klima“, wird das Konzeptnetz auf die deutsche Seite hinübergezogen.

Fazit

Die Mischung der Textsammlungen aus unterschiedlichen Sprachen kann den Speicherplatz um ca. 15% reduzieren. Eine Nebenwirkung stellen aber die transliterierten Wörter dar. Obwohl das Problem der Nebenwirkung hier nicht sehr stark auftritt, kann nicht garantiert werden, dass sie den Nutzer dabei nicht behindert, das Konzeptnetz zu verstehen, falls die Begriffe aus verschiedenen Sprachen gemischt werden.

Der Grund ist, dass die Assoziationsstärke der deutschen Begriffe höher ist als die der englischen. Diese Untersuchung verrät uns, dass die bilinguale Suche durch das Konzeptnetz auch funktionieren sollte, wenn der Zielcontainer zufällig andere fremde Texte enthält. Solange es kein transliteriertes Wort in dem fremden Text und als Anfrage gibt, ergibt sich ein Ergebnis ohne fremdes Dokument.

5.3 Konzeptnetzänderung

Diese Untersuchung geht um Folgendes,

- Wie wird das Konzeptnetz bei Größenänderung des Containers verändert?
- Ist die Entwicklung des Konzeptnetzes bezüglich des deutschen Containers ähnlich zu der Entwicklung des Konzeptnetzes bezüglich des englischen Containers?
- Ergibt sich durch die Vergrößerung des Containers eine Stabilisierung des Konzeptnetz?

Experiment

Die Wortliste, die das Konzeptnetz bildet, wird betrachtet. Sie wird bezüglich der Assoziationsstärke geordnet. Die Betrachtung auf der Wortliste wird in fünf Blöcke unterteilt, von Rangplatz 1 bis 10, von 11 bis 20, von 21 bis 30, von 31 bis 40 und von 41 bis 50. Die Erweiterung des Containers beginnt mit ein zu zwei Jahren, dann von zwei zu drei Jahren und schließlich von drei zu vier Jahren. Die Anfragen werden durchgeführt, um die verschiedenen Wortlisten zu bilden.

Rangplatz	**De-Anfrage 1**																	
	1J → 2J						2J → 3J						3J → 4J					
	1	2	3	4	5	x	1	2	3	4	5	x	1	2	3	4	5	x
1 – 10	6	2	0	0	1	1	6	3	0	0	0	1	8	0	1	0	0	1
11 – 20	1	0	1	1	0	7	4	3	0	1	1	1	1	6	2	0	0	1
21 – 30	2	3	0	2	0	3	0	2	6	1	0	1	1	1	4	0	1	3
31 – 40	1	1	2	1	0	5	0	1	4	2	0	3	0	2	2	3	3	0
41 – 50	0	1	2	1	1	5	0	1	0	1	1	7	0	0	0	4	1	5
Summe	10	7	5	5	2	21	10	10	10	5	2	13	1	9	9	7	5	10

Tabelle 10 Die „1“ auf der Spalte entspricht dem Block „1 bis 10“ aus dem vorhergehenden Jahr. Das „x“ repräsentiert die Anzahl der neuen Terme. Die Anzahl der gefundenen Wörter wird für jeden Block zusammengezählt.

Die Änderung der Wortliste auf den ersten n-zehnten Plätzen wird beobachtet, man findet z.B. auf den ersten dreißig Rangplätzen dieselben 15 Begriffe bei der Erweiterung von einem Jahr zu zwei Jahren „6+2+0“ im ersten Block und „1+0+1“ im zweiten Block sowie „2+3+0“ im dritten Block. Die Zahlen kennzeichnen hier die Anzahl der unterschiedlichen Begriffe. Begriffe, die auf hinteren Rangplätzen liegen, werden hier vernachlässig. Das folgende Beispiel stellt die absolute Anzahl und die Ähnlichkeit in Prozent entsprechend der deutschen Anfrage 1 dar.

Rangplätze	**De-Anfrage 1**					
	1j → 2j	%	2j → 3j	%	3j → 4j	%
Erste 10.	6	0,60	6	0,60	8	0,80
Erste 20.	9	0,45	16	0,80	15	0,75
Erste 30.	15	0,50	24	0,80	24	0,80
Erste 40.	23	0,58	33	0,83	31	0,78
Erste 50.	29	0,58	37	0,74	40	0,80

Tabelle 11 Die Ähnlichkeit der Begriffe gemäß der Containervergrößerung in absoluter Anzahl und in Prozent verteilen sich auf den n-zehnten Plätzen.

Die Ähnlichkeit in Prozent auf den unterschiedlichen ersten Rangplätzen unter der selben Containererweiterung ist relativ gleich. Die Änderungen der jeweiligen n-zehnten Rangplätze sind unwesentlich verschieden. Die Ergebnisse gemäß einer anderen Anfrage ergeben sich genauso wie bei der deutschen Anfrage 1. Außerdem wird die Prozentzahl von den Begriffen der jeweiligen n-zehnten Rangplätze durch die Vergrößerung des Containers gesenkt.

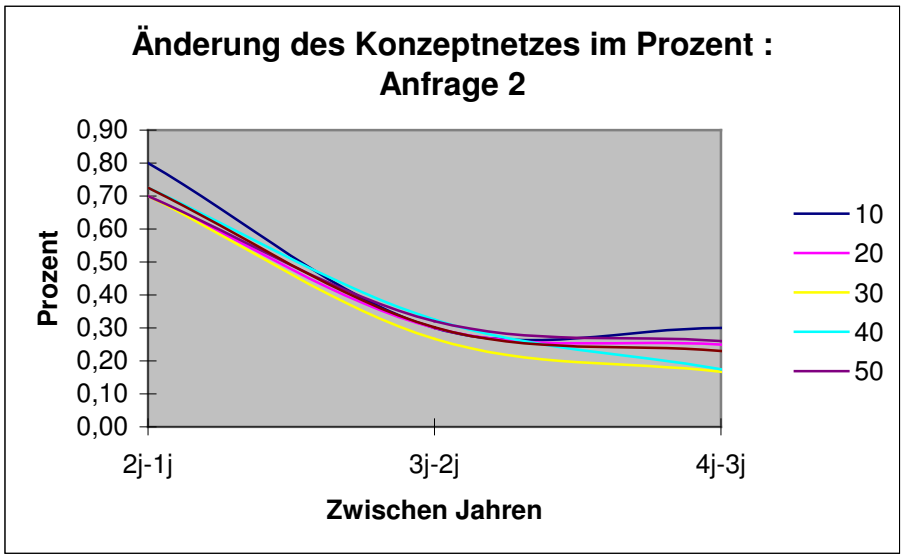


Abbildung 24 Die Tendenz der Begriffe-Änderung entsprechend der deutschen Anfrage 2.

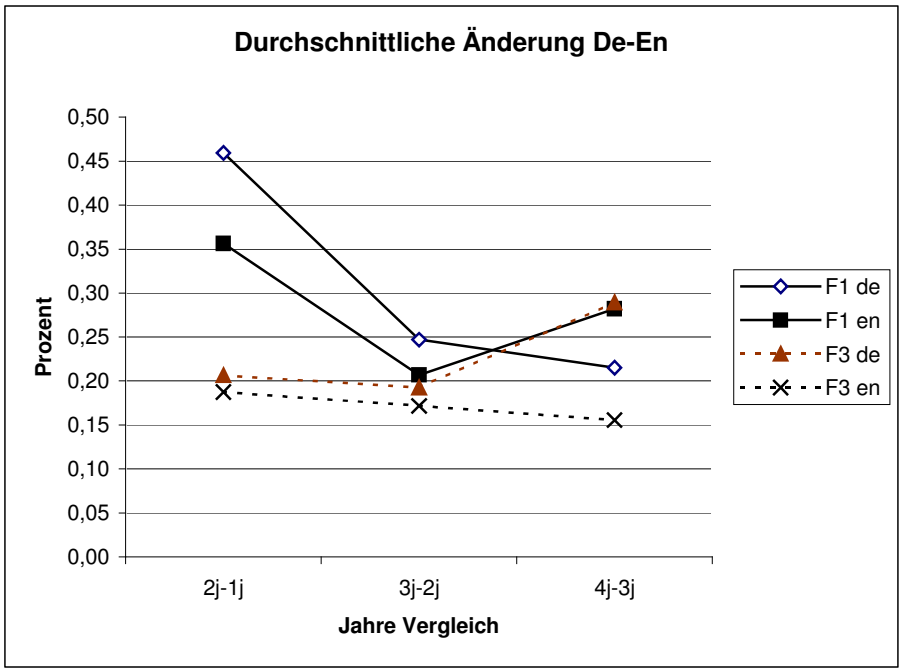


Abbildung 25 Durchschnittliche Begriffänderung gemäß der deutschen Anfragen 1 bzw. 3 und der englischen Anfragen 1 bzw. 3.

Obwohl die Vergrößerungen des Containers im Verhältnis „1:1“, „2:1“ und „3:1“ stattfinden, ist die prozentuale Änderung unabhängig davon. Als Beweis sei hier die prozentuale Änderung von einem Jahr zu zwei Jahren im Vergleich zu von zwei Jahren zu vier Jahren genannt (siehe Tabelle 12).

Rangplätze	**Anfrage 1**		**Anfrage 2**		**Anfrage 3**		**Anfrage 4**	
	1J→2J	2J→4J	1J→2J	2J→4J	1J→2J	2J→4J	1J→2J	2J→4J
Erste 10.	0,40	0,40	0,80	0,40	0,00	0,40	0,50	0,20
Erste 20.	0,55	0,35	0,70	0,45	0,30	0,40	0,45	0,25
Erste 30.	0,50	0,33	0,70	0,40	0,27	0,33	0,47	0,30
Erste 40.	0,43	0,28	0,73	0,43	0,23	0,25	0,50	0,25
Erste 50.	0,42	0,36	0,70	0,44	0,24	0,26	0,54	0,22
Durchschnitt	0,46	0,34	0,73	0,42	0,21	0,33	0,49	0,24

Tabelle 12 Die Begriffe-Änderung in Prozent bei doppelter Containervergrößerung.

Dieser Vergleich zeigt ganz klar, dass die Begriffe-Änderung bei Containervergrößerung von zwei zu vier Jahren fast halb so groß ist wie die Begriffe-Änderung der Containervergrößerung von einem zu zwei Jahren, obwohl doppelt so stark vergrößert wurde. Nur bei der Anfrage 3 ist es signifikant. Der Grund dafür ist, dass die Begriffe auf der Wortliste gemäß der Anfrage 3 die Plätze wechseln.

Der Platzwechsel wurde bisher noch nicht berücksichtigt. Einige Begriffe wechseln nur von einem Block zu einem anderen Block in der Nähe. Die Begriffe werden auf jedem Block auf der Wortliste betrachtet und in vier Gruppen aufgeteilt im Sinne, „Aufstieg“, „Bleiben“, „Senkung“ und „Verlust“. „Aufstieg“ bedeutet hier, dass die Begriffe von einem unteren Block zu einem höheren Block aufsteigen, während „Senkung“ das Gegenteil darstellt. „Bleiben“ bezeichnet die in dem gleichen Block verbleibenden Begriffe und „Verlust“ repräsentiert die Begriffe aus den ersten 50ten-Rangplätzen, die verloren gehen. In Tabelle 13 werden die Platzwechsel und die Verluste in Prozent bei den vier deutschen Anfragen dargestellt.

%	Frage 1			Frage 2			Frage 3			Frage 4		
	1j → 2j	2j → 3j	3j → 4j	1j → 2j	2j → 3j	3j → 4j	1j → 2j	2j → 3j	3j → 4j	1j → 2j	2j → 3j	3j → 4j
Aufstieg	14%	12%	14%	6%	14%	24%	14%	14%	14%	18%	12%	14%
Bleiben	16%	36%	44%	6%	40%	32%	44%	46%	36%	18%	56%	42%
Senkung	28%	26%	22%	8%	14%	18%	18%	18%	22%	10%	14%	26%
Verlust	42%	26%	20%	70%	32%	26%	24%	22%	28%	54%	18%	18%

Tabelle 13 Die Platzwechsel und die Verluste an Begriffen in Prozent.

Aus der Tabelle 13 kann man ersehen, dass abgesehen von dem Verlust die meisten Begriffe in demselben Block verbleiben. Obwohl manche Begriffe aufgestiegen oder

gesunken sind, wechseln sie meistens aber nur um einen Block. Die Tabelle 14 zeigt die Prozentzahl der relativen Platzwechsel, die bei „Aufstieg“ und „Senkung“ lediglich um einen Block geschehen.

%	Frage 1			Frage 2			Frage 3			Frage 4		
	1j → 2j	2j → 3j	3j → 4j	1j → 2j	2j → 3j	3j → 4j	1j → 2j	2j → 3j	3j → 4j	1j → 2j	2j → 3j	3j → 4j
Aufstieg	4%	4%	4%	4%	0%	2%	4%	0%	8%	2%	6%	6%
Bleiben	40%	66%	70%	18%	64%	68%	68%	78%	58%	38%	72%	76%
Senkung	14%	4%	6%	8%	4%	4%	4%	0%	6%	6%	4%	0%
Verlust	42%	26%	20%	70%	32%	26%	24%	22%	28%	54%	18%	18%

Tabelle 14 Relativer Platzwechsel und der Verlust an Begriffe in Prozent.

Aus der Tabelle 14 wird deutlich, dass die Platzwechsel bei Vergrößerung des Containers fast nur in naher Umgebung liegen. Diese Änderung ist nicht so dramatisch. Die Begriffe bleiben meist im Konzeptnetz. Außerdem wird die Prozentzahl der relativen verbleibenden Begriffe aus der Tabelle 14 wegen der Containervergrößerung größer. Dies gilt ebenfalls für den englischen Container. Somit könnte man schon hier von einer Stabilisierung des Konzepts durch doppelte Containervergrößerung sprechen.

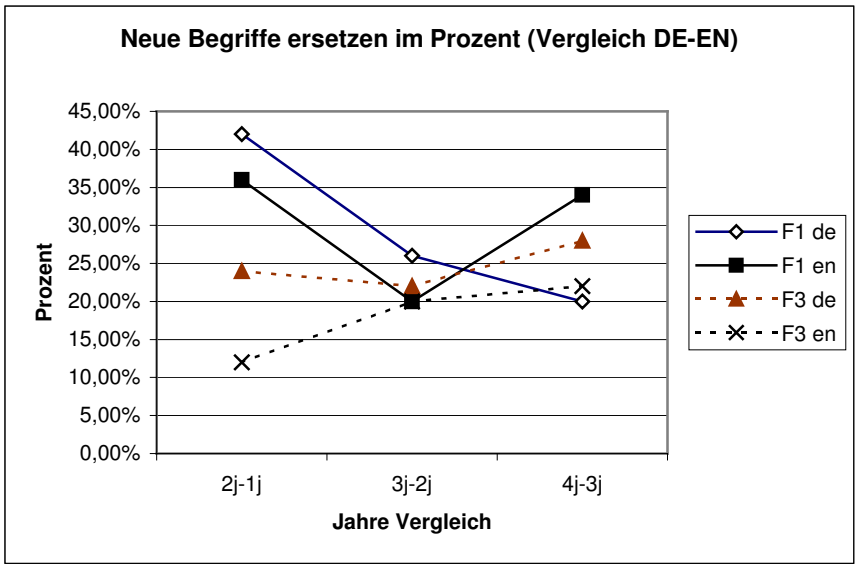


Abbildung 26 Die Prozentzahl neuer Begriffe bei Containervergrößerung gemäß der deutschen und englischen Anfrage 1 und Anfrage 3.

Die neu gewonnenen Terme bei Vergrößerung des Containers in Ein-Jahres-Schritten der jeweiligen n-zehnten Rangplätze werden gezählt, um das Änderungsverhalten zu beobachten. Zwei Bereiche werden betrachtet, die ersten 30 Plätze und dann die ersten

50 Plätze. Die neuen Begriffe, die in dem Bereich wegen der Vergrößerung des Containers vorkommen, werden gezählt. Als Ergebnis zeigt sich, dass sich die Stabilität der Liste erst bei den hohen Rangplätzen ergibt (siehe Tabelle 15).

		Deutsch		Englisch	
		Erste 30	Erste 50	Erste 30	Erste 50
Anfrage 1	1j → 2j	37%	42%	20%	34%
	2j → 3j	10%	26%	10%	20%
	3j → 4j	16%	20%	23%	36%
Anfrage 3	1j → 2j	13%	24%	10%	22%
	2j → 3j	10%	22%	10%	20%
	3j → 4j	13%	28%	0%	12%

Tabelle 15 Die neuen Begriffe auf den ersten 30sten-Rangplätzen und den ersten 50sten-Rangplätzen entsprechend den deutschen Anfragen und den übersetzten englischen Anfragen in Prozent.

Ein Verlust an alten Begriffen ist manchmal auf einem hohen Rangplatz zu beobachten, z.B. bei der deutschen Anfrage 4 werden einige Begriffe aus der Wortliste des kleineren Containers bei der Wortliste des größeren Containers deutlich nach unten verschoben. Ursache ist der Einfluss der Zusatztexte. Obwohl die Anzahl der relevanten Dokumente bei der Containervergrößerung nicht erhöht wird, kann das Konzeptnetz von neuen Assoziationen aus der Zusatzsammlung beeinflusst werden.

Rangplatz	De Anfrage 4																	
	1J → 2J						2J → 3J						3J → 4J					
	1	2	3	4	5	x	1	2	3	4	5	x	1	2	3	4	5	x
1 – 10	5	3	0	0	0	2	8	1	0	0	0	1	8	0	1	0	0	1
11 – 20	1	2	2	1	0	4	0	6	0	1	2	1	1	7	0	1	0	1
21 – 30	1	1	1	2	0	5	0	1	7	1	0	1	0	2	4	2	1	1
31 – 40	1	0	0	0	1	8	0	1	2	4	1	2	0	0	5	2	2	1
41 – 50	0	0	1	0	1	8	0	0	1	2	3	4	0	0	0	5	0	5
Summe	8	6	4	3	2	27	8	9	10	8	6	9	9	9	10	10	3	9

Tabelle 16 Die Anzahl der Begriffänderungen auf Wortliste gemäß der deutschen Anfrage 4. Die Summe zeigt die Anzahl der verbliebenen Begriffe aus dem kleineren Container je nach Block an.

Im Ein-Jahres-Container gibt es zwei 100%-Dokumente gemäß der deutschen Anfrage 4. Im Zwei-, Drei- und Vier-Jahres-Container befinden sich dagegen fünf 100%-Dokumente. Wie man sehen kann, kommen zwischen dem Ein-Jahres-Container und dem Zwei-Jahres-Container drei 100%-Dokumente hinzu, während im Folgenden keine weiteren 100%-Dokumente hinzukommen. Bemerkenswert dabei ist, dass zwar wie er-

wartet bei dem Hinzukommen von drei weiteren 100%-Dokumenten sich Veränderungen in der Wortliste ergeben, aber auch Veränderungen zu beobachten sind, wenn keine weiteren 100%-Dokumente mehr hinzukommen. Ein Beispiel dafür ist, dass bei der Containeränderung „3J → 4J“ in den ersten zehn Rängen (Spalte 1) ein und in den Rängen 11 bis 20 (Spalte 2) ein weiterer Begriff verloren geht. Allerdings ist eine Veränderung am stärksten auf den hinteren Rangplätzen (z.B. in Spalte 5) zu erwarten.

Fazit

Die Anzahl alter Begriffe bei der Vergrößerung des Containers ist deutlich höher, wenn die neue Textsammlung mit dem vorhergehenden Container zusammen gebildet wurde. Dies hängt nicht davon ab, ob die erweiterte Sammlung genau so groß oder kleiner ist als im früheren Container. Jeder n-zehnte Rangplatz der Wortliste hat in Prozent der gebliebenen Begriffe einen nahezu gleichen Wert. Dadurch ergibt sich kein Unterschied im Konzeptnetz, egal ob das Netz klein oder groß ist.

Das Änderungsverhalten des deutschen und englischen Konzeptnetzes ist ähnlich, weil der inhaltliche Prozess unabhängig von der Sprache ist. Daher ist bei dieser Untersuchung völlig irrelevant, ob man auf dem englischen oder deutschen Korpus sucht.

Wenn eine neue Textsammlung in den Container eingefügt wird, verändert sich die Wortliste hauptsächlich auf den hinteren Rangplätzen. Die neuen Begriffe tauchen in dem Konzeptnetz abhängig davon auf, wie viele Begriffe auf dem Konzeptnetz vom Nutzer eingestellt wurden. Weil das Konzeptnetz von den Termen in der Anfrage abhängig ist, kann man nicht feststellen, wann die Netze in den entgültigen Zustand übergehen. Aber es scheint, dass sich das Konzeptnetz bei stetiger Vergrößerung des Containers langsam entwickelt und stabilisiert.

5.4 Suche im nicht-parallelen Korpus

Der nicht-parallele Korpus wird hier in zwei Container aufgeteilt, um die Suche mittels Konzeptnetz in der nicht-parallelen Situation zu testen. Der Inhalt der getroffenen Do-

kumente wird manuell verglichen. Die aktivierten Konzeptgliederungen beider Container werden ebenfalls daraufhin beobachtet, ob es trotz Nichtparallelität noch Beziehungen zwischen ihnen gibt.

Experiment

Der nicht-parallele Korpus wird aus dem parallelen Korpus konstruiert, indem gerade bzw. ungerade Dateinummern gewählt werden (siehe Tabelle 17). Die Anfragen werden vorbereitet, um die Suche zu testen. Die deutschen und englischen Konzeptnetze und die getroffenen Dateien werden verglichen. Anzumerken ist, dass die Anzahl der Dateien zwischen Englisch und Deutsch nicht identisch ist, weil sie sich nach der Marke „Chapter ID“ auftrennen lassen und die Anzahl der deutschen und englischen Marken in den parallelen Korpora nicht gleich ist. Es gibt aber nach der nicht-parallelen Auftrennung zufällig 10 parallele Dokumente. Die so gebildeten Container werden je nach Sprache indexiert.

	1998	1999	2000	2001	Anzahl der Dateien
Deutsch	ungerade	gerade	ungerade	gerade	1.180
Englisch	gerade	ungerade	gerade	ungerade	1.362

Tabelle 17 Ausgewählte Dateien für die deutsch-englisch nicht-parallele Korpora.

Für die deutsche Anfrage „Gentechnik, Lebensmittel, Lebensmittelsicherheit, GVO“ entsprechend der englischen Anfrage „genetic engineering, food, food safety, GMOs“ ergeben sich zwei deutsche bzw. fünf englische 100%-Dokumente. Alle getroffenen Dokumente drehen sich um das Thema „die Sicherheit des Lebensmittels durch genetisch veränderte Organismen“ (siehe Tabelle 18). Wenn man beide Konzeptnetze betrachtet, sieht man, dass viele Begriffe die Übersetzungspaare auf der anderen Seite ergeben können bzw. semantische Vergleichbarkeiten besitzen. Die semantische Vergleichbarkeit ist somit entweder abhängig von der unterschiedlichen Nutzungsweise der jeweiligen Sprache oder der Schreibweise unterschiedlicher Autoren. Anhand des deutschen Konzeptnetzes, das für 50 Begriffe erstellt wurde, kann man erkennen, dass vier Begriffe Suchwörter sind und ca. 50% der restlichen Begriffe einen ähnlichen semanti-

schen Sinn haben. Weil das Konzeptnetz bei einer globalen Analyse aller Schlüsselwörter im Korpus ermittelt wird, hängen die nicht nur zum relevanten Dokument gehörenden restlichen Begriffe oft am Netz, wie man auch an den 50% restlichen Begriffen aus dem obigen Beispiel sehen kann, die keinen semantisch ähnlichen Sinn haben

	Dokument	Thema/Umgebung
DE	ep-01-02-13.txt2	*Freisetzung genetisch veränderter Organismen* / abschließende Erklärung der Richtlinien über GVO; Maßnahme in Bezug auf Haftung, Rückverfolgbarkeit und Kennzeichnung; einige Diskussionen über Medizin, kommerzielle Zwecke, Lebensmittelsicherheit.
	ep-01-03-15.txt4	*Abstimmung* / Anteil des Dokuments – Gentechnik in der Medizin, die Biotechnologie in der Landwirtschaft. Lebensmittelsicherheit mit der Zulassung der GVO. Das Beispiel von der BSE-Krankheit.
EN	ep-00-04-11.txt10	*Deliberate release into the environment of GMOs* / dangers of genetically modified organisms; GMO products throughout Europe in a safe; stringent standards; approval and control of GMO crops and food; govern monitoring, labeling and informing the public; pors and cons of GMOs; health safety and environment protection.
	ep-00-03-14.txt4	*Cocoa and chocolate products* / using GMOs in industrial processes, which make it possible to obtain cocoa-butter equivalents; foodstuff; food safety
	ep-01-03-15.txt3	*Biotechnology industry* / production of food; safety, when it comes to consumers, the work of farmers and the whole food processing chain; agricultural industries; food form developing countries (emphasis on GMO in food and agricultural products).
	ep-00-10-25.txt2	*Food safety* / food safety using genetic engineering
	ep-00-03-15.txt4	*Vote* / some part of text are talked about cocoa, food safety, genetic engineering and GMOs in chocolate.

Tabelle 18 die Themen und Inhalte der relevanten Dokumente entsprechend der deutschen Anfrage „Gentechnik, Lebensmittel, Lebensmittelsicherheit, GVO“ bzw. der englischen Anfrage „genetic engineering, food, food safety, GMOs“.

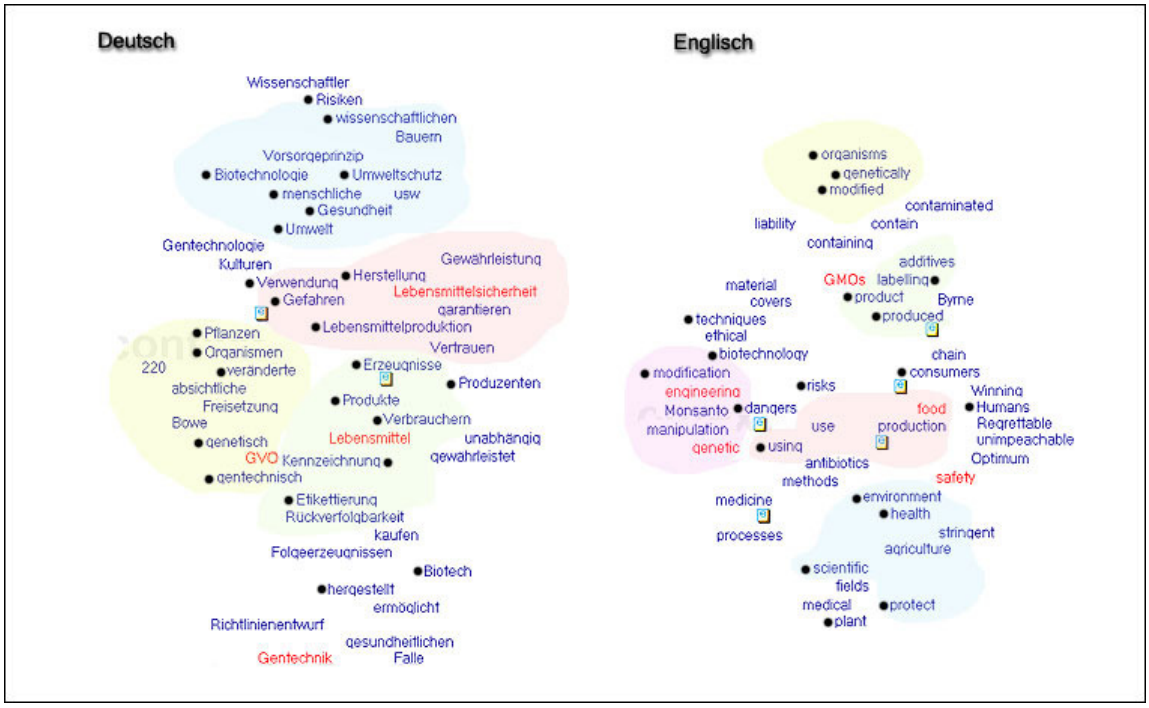


Abbildung 27 Konzeptnetze entsprechend der Anfrage „Gentechnik, Lebensmittel, Lebensmittelsicherheit, GVO“ bzw. „genetic engineering, food, food safety, GMOs“. Die schwarzen Punkte repräsentieren die Übersetzungspaare und die semantisch vergleichbaren Paare. Die farbigen Bereiche sind quasi einzelne abgebildete Konzepte.

Obwohl die letzte Suche als erfolgreich erscheint, bedeutet dies nicht, dass die 100-prozentig relevanten Dokumente die richtige Antwort repräsentieren. Bei der Suche mit der deutschen Anfrage „Konflikt, Region, Waffenherstellern“ findet sich das einzige 100-prozentig relevante Dokument, in dem es um die Konfliktverhütung und die Friedenserhaltung geht. Bei der englischen Suche hingegen führt das einzige 100-prozentig relevante Dokument gemäß der Anfrage „conflict, region, armament manufacturers“ nicht zu demselben Thema, weil es um viele kurze Themen oder Unterthemen geht. Das Wort „conflict“ und das Wort „region“, das sich aber auch in „Humanitarian aid in the Caucasus“ findet, findet man in „Chechnya conflict“ und „Humanitarian aid for children“. Leider taucht das Wort „manufacturers“ mit dem Wort „car“ in „Pedestrian-friendly car“ und das Wort „armament“ mit dem Wort „policy“ in „Defense industries“ auf. Grund dafür ist, dass es sich in dem getroffenen englischen Dokument um das Thema „Question Time“ dreht. Dadurch gehört vieles davon zu diesem Hauptthema.

Die Übernahme eines Schlüsselwortes aus dem relevanten Dokument ist eine gute Möglichkeit, um auf ein Unterthema einengen zu können. Das dritte Beispiel zeigt, dass diese Methode ein gutes Ergebnis liefern kann. Die anfängliche deutsche Anfrage „Landwirtschaft, Umweltschutz, Agrarpolitik, Umwelteffekt“ wird zunächst abgearbeitet. Das einzige 100-prozentig relevante Dokument „ep-98-10-22.txt5“ mit dem Thema „Umwelt und landwirtschaftliche Produktion – Beihilfe für Aufforderungsmaßnahmen – Bergregion“ wird ermittelt. Die englische Anfrage „environment protection, evironment effect, agriculture, agriculture policy“ wird mit den zusätzlichen übersetzten Wörtern „aid, afforestation“ aus dem Thema und dem Inhalt des deutschen Dokumentes „Beihilfe zur Aufforstung“ aufgefüllt und abgearbeitet. Die 100-prozentig relevanten Dokumente werden in der Tabelle 19 beschrieben.

	Dokument	Thema/Umgebung
DE	ep-98-10-22.txt5	*Umwelt und landwirtschaftliche Produktion – Beihilfe für Aufforstungsmaßnahmen – Bergregion*
EN	ep-98-10-23.txt2	*Vote* / some part of text are talked about environment protection to promote the agricultural products, forest aid program, agriculture policy, aid for afforestation.
	ep-01-05-15.txt11	*Question Time* / under subject: Fires and reafforestation in Greece found many searching words but the word “aid” was found under other irrelevant subjects.
	ep-98-06-16.txt6	*Reform of CAP (Agenda 2000)* / agriculture policy, aid to forestly as agriculture, protection of environment and good animal welfare, agriculture model, aid for afforestation.

Tabelle 19 Die Themen und Inhalte der relevanten Dokumente entsprechend der deutschen Anfrage „Landwirtschaft, Umweltschutz, Agrarpolitik, Umwelteffekt“ bzw. der englischen Anfrage „environment protection, evironment effect, agriculture, agriculture policy, aid, afforestation“.

Wie bei der ersten Anfrage sind die Begriffe auf den Konzeptnetzen fast zu 50% gleich. Dieses ergibt sich aber nicht bei dem zweiten Beispiel oder bei der nächsten Anfrage.

Der vierte Versuch ist erfolglos. Auf dem deutschen Konzeptnetz steht das Wort „Haushalte“ ganz nah an den anfänglichen Suchwörtern „Ölkrise“ und „Wirtschaftspolitik“. Es gibt wahrscheinlich einen Zusammenhang. Ohne die Übersetzung des Wortes

„Haushalte“ hinzuzufügen werden zahlreiche 100%- Dokumente durch die anfängliche englische Anfrage „oil crisis, economic policy“ geliefert. Mit „households“, werden die sieben englischen 100%-Dokumente gewonnen. Die meisten von diesen sind aber über „Vote“ und „Question Time“, die sich allerdings um viele unterschiedliche Teilthemen dreht. Obwohl zwei deutsche Dokumente der Anfrage entsprechend sich um die gewünschten Themen drehen, sind die ermittelten englischen Dokumente leider nicht zutreffend.

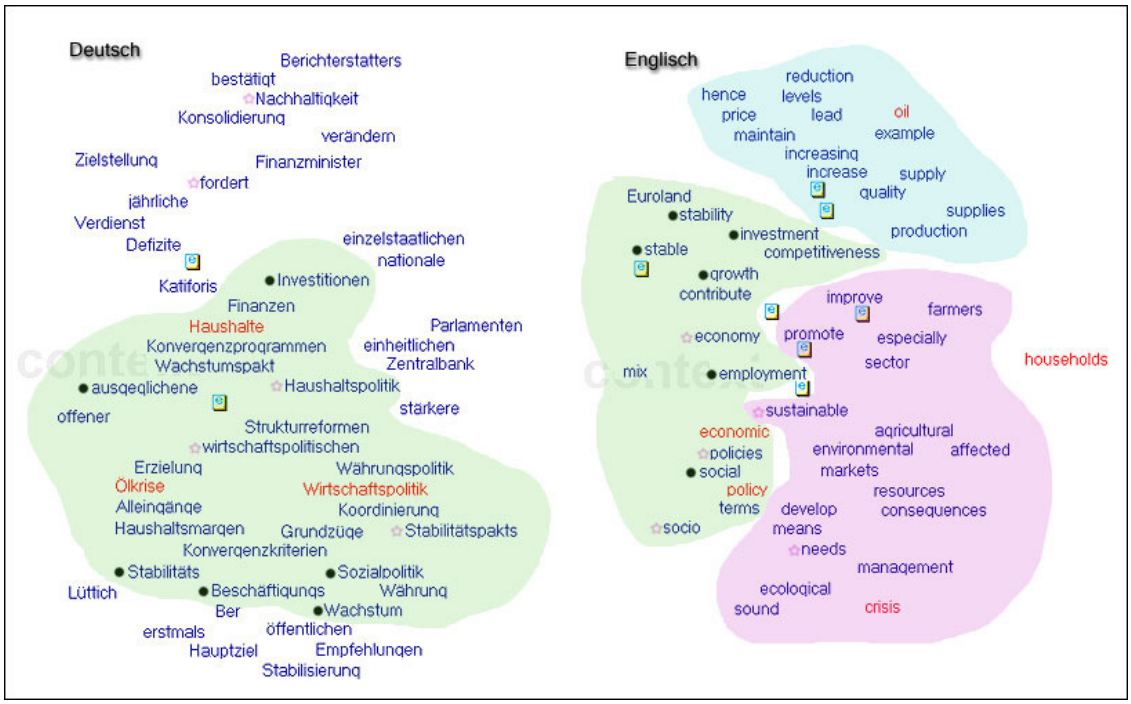


Abbildung 28 Die Konzeptnetze entsprechend der Anfrage „Ölkrise, Wirtschaftspolitik, Haushalte“ bzw. „oil crisis, economic policy, households”. Die schwarzen Punkte repräsentieren die Übersetzungspaare und die Sternzeichen bezeichnen die semantisch vergleichbaren Paare. Die farbigen Bereiche sind quasi einzelne abgebildete Konzepte.

Auf den Konzeptnetzen ist zu erkennen, dass ein großer Teil des englischen Konzeptnetzes drei Konzepte besitzt. Das Konzept unten rechts kann „Umwelt, Landwirtschaft, Wirtschaft und Nebenwirkung der Krankheit“ bilden. Das obige Konzept ist quasi das Thema „Ölpreis, Ölqualität usw.“. Das linke Konzept entspricht wahrscheinlich „Wirtschaftwachstum und Sozialpolitik“, während es sich bei dem deutschen Konzeptnetz größtenteils um ein dem linken ähnlichen Thema handelt. Das Wort „household“ hat ei-

nen ziemlich schwachen Zusammenhang mit der anfänglichen englischen Anfrage. Darüber hinaus kann ein Hinzufügen von „household“ in der englischen Anfrage das Ergebnis ablenken.

Fazit

Wenn die Konzeptnetze beider Sprachen viele semantik- bzw. übersetzungsvergleichbare Begriffe haben, führen sie eventuell zu entsprechenden bilingualen Dokumentenpaaren hin. Die „Ähnlichkeit des Konzeptnetzes“ kann aber auch die Vergleichbarkeit des konzeptionellen Anteils der Netze meinen.

Der Unterschied auf den Konzeptnetzen stammt nicht nur aus den Beziehungen zwischen den Suchwörtern und den Begriffen, sondern auch aus der unterschiedlichen sprachlichen Nutzungsweise und der autorspezifischen Schreibweise. Durch die sprachliche Vorverarbeitung kann das Problem der unterschiedlichen sprachlichen Nutzungsweise teilweise verhindert werden, indem die Stammform und das Vernachlässigen von unbenötigten Wortarten die Ablenkung durch ungeeignete Begriffe verhindern kann. Als Konsequenz davon ergibt sich auch eine Änderung der Assoziationsstärke. Dies könnte die sprachliche Symmetrie bringen, die für den Vergleich der Konzeptnetze nötig ist.

Die Nutzung der zusätzlichen Begriffe aus den relevanten Dokumenten der Ausgangsprache als zusätzliche Übertragungsbegriffe ist eine sinnvolle Methode. Dies kann manuell sowie automatisch erfolgen. Obwohl die manuelle Auswahl einfach und direkt ist, muss der Nutzer viel Zeit aufwenden, um die getroffenen Dokumente durchzulesen. Bei der automatischen Auswahl können die ersten *n* gewonnenen Begriffe der Worthäufigkeit oder der Assoziationsstärke nach gemäß der Suchanfrage übernommen werden. Diese Methode kann man „Pseudo-Relevant-Feedback“ nennen.

Obwohl die TrefferDoc-Funktion der SENTRAX die tolerante Suche mittels der SpaCAM-Technologie ermöglicht, können die getroffenen Dokumente eventuell von dem gesuchten Thema abweichen. Die SENTRAX ist eine Volltextsuche mittels Musterabgleichung. Wenn ein Dokument viele unabhängige Themen besitzt, kann die Suche ab-

gelenkt werden. Bei solchen Dokumenten können einige Suchbegriffe in einem Thema und andere in einem anderen Thema vorkommen. Auch wenn die suchenden Begriffe in einem Dokument weit von einander getrennt stehen, wertet die Musterabgleichung diese bei der Volltextsuche als relevant. Aufgrund der Vielfältigkeit der Themen und der schwachen Beziehungen durch weit auseinander stehende Wörter soll ein solches Dokument als unrelevant angesehen werden.

6 ZUSAMMENFASSUNG UND AUSBLICK

6.1 Schlussfolgerungen

1. Die Konzeptnetze funktionieren nicht nur als Hilfswerkzeuge, um im parallelen Korpus die parallelen relevanten Dokumente abzurufen, sondern ermöglichen auch einen Lernprozess während der Suchschritte. Der Benutzer erkennt, zu welchen Themen das Konzeptnetz hinführen kann, wovon die Begriffe des Netzes handeln könnten und ob die Suche auf dem richtigen Weg ist usw. Dies ist aber nützlich sowohl für den Anfänger als auch für den Experten.
2. Die Auswahl verwandter Begriffe bei den Suchanfragen verstärkt das Suchkonzept. Die direkt auf die relevanten Dokumente bezogenen zusätzlichen Begriffe haben keine Wirkung auf die Änderung des Treffers, aber ihre Übersetzungen können den Rangplatz der völlig unverwandten Dokumente in der Zielsprache verkleinern. Somit kann das Suchthema verschärft werden.
3. Mit Hilfe der Kontext-Umgebung kann die Suche besser gesteuert werden, indem suchverwandte Begriffe zusätzlich ausgewählt werden. Dies hilft besonders bei der manuellen bilingualen Suche, wo sich dadurch dasselbe Ziel besser einschränken lässt.
4. Wenn man gewünschte Dokumente in der Ausgangsprache gefunden hat, kann man die SimilarDoc-Funktion einsetzen, um zum ausgewählten Dokument weitere relevante zu finden. Wenn man eine parallele Dokumentensammlung hat, werden so fast alle zu einem festen Dokument ähnlichen auch auf der parallelen Seite geliefert.
5. Die deutschen Komposita haben nun Korrespondenzen zu der englischen Nominalphrase. Weil ein deutsches Kompositum hier eine englische Wortgruppe dort repräsentiert, wird der Prozess des jeweiligen Assoziierens unterschiedlich sein. Diese Unsymmetrie verursacht einen Unterschied in der Worthäufigkeit sowie in der direkten Assoziation (vgl. Abschnitt 4.1.1). Im deutschen Text findet man

z.B. das Wort „Umwelt“ und „Umweltschutz“, während im englischen Text Worte wie „environment“ und „environment protection“ bzw. „environmental protection“ stehen. Bei diesem Beispiel wird die Häufigkeit des Wortes „Umwelt“ bzw. „Umweltschutz“ jeweils nur eins sein. Im englischen Text hingegen sind die Häufigkeiten für das Wort „environment“ zwei und für „protection“ eins bzw. für „environment“, „environmental“ und für „protection“ jeweils eins, je nachdem, welcher Ausdruck sich im Text befindet. Dies wird nun aber weniger bedeutsam, weil zu einer Anfrage an die anderssprachige Datenbasis, die konventionell aus einer Übersetzung der Begriffe aus der Anfrage in dieser Sprache besteht, nun die Übersetzung weiterer Begriffe aus dem Konzeptnetz hinzukommen. Somit werden die Suchkonzepte der beiden Suchen ausgeglichen.

6. Die TrefferDoc-Funktion und ContextMap-Funktion sind unabhängig voneinander, sie beruhen auf unterschiedlichen Musterabgleichungstechniken. Die ContextMap-Funktion entsteht aus den Assoziationen zwischen den Begriffen (in der Datenbasis), wohingegen die TrefferDoc-Funktion von der einfachen Musterabgleichung mithilfe der SpaCAM-Technologie erzeugt wird. Diese Abgleichung enthält keine Relationen oder Assoziationen zwischen den Suchwörtern. Das bringt es mit sich, dass sich manchmal Dokumente in der 100-Prozent Zone der Trefferliste finden (weil die Eingabebegriffe enthalten sind), die aber inhaltlich nicht zwingend relevant sein müssen.
7. Die Übersetzung ohne die konzeptionelle Semantik verursacht die Mehrdeutigkeit, die normalerweise das Hauptproblem der krosslingualen Suche ist. Durch den Einfluss der zusätzlichen umgebungsbezogenen Begriffe kann sich die Übersetzung deutlicher und enger, gezielter auswirken.
8. Obwohl manche deutschen Komposita keine englische Übersetzung im (elektronischen) Wörterbuch haben, können die geeigneten zusätzlichen Begriffe aus dem Konzeptnetz die Lücke füllen. Trotzdem bleibt dies häufig noch ein Problem.

9. Die bilinguale Suche funktioniert symmetrisch, also von D→E als auch von E→D. Das liegt an der Gleichartigkeit der Methoden, erfordert jedoch im gegenwärtigen Stadium noch Wörterbücher gleichwertiger Art.
10. Der Unterschied der Größe zwischen Ausgangs- und Zielcontainer hat keinen Einfluss auf die bilinguale Suche. Es müssen unter Umständen wegen der Vergrößerung des Containers aber mehr Begriffe ausgewählt werden, um das Suchergebnis besser beschränken zu können. Bei kleinerem Zielcontainer ist ein Zusatzbegriff oft nicht erforderlich.
11. Wenn der Zielcontainer kein relevantes Dokument enthält, wird der Zusammenhang zwischen Konzeptnetzen, denen dieselbe Suchanfrage zu Grunde liegt, praktisch nicht gefunden.
12. Die Entfernung einiger Dokumenten aus dem Container hat aber keine Auswirkung auf die abgerufene Trefferliste. Die nachkommenden Dokumente setzen sich auf die Plätze nacheinander in der Reihefolge, wo die entfernten vorher waren.
13. Die Veränderung der Konzeptnetzansicht durch die Abwesenheit einiger relevanten Dokumente kann sich deutlich auswirken, besonders dann, wenn der Container nicht viele weitere relevante Dokumente gemäß der Anfrage enthält.
14. Mehrsprachliche Textsammlungen können in einen einzigen Container gepackt werden. Das stört kaum; es kann aber eine Verzerrung der Konzeptnetze auftreten, weil der Container nun die transliterierten Wörter beinhaltet und mehrsprachliche Begriffe im selben Netz vermischt werden. Ansonsten funktioniert die Suche mittels der Konzeptnetze entsprechend der in der gewählten Sprache eingegebener Anfrage einwandfrei.
15. Das Konzeptnetz hat die Tendenz zur Stabilität, wenn der Container groß genug gewählt war. Das bedeutet auch, dass das Konzeptnetz fast unverändert bleibt, wenn man einem großen Container eine neue Textsammlung hinzufügt.
16. Der Beständigkeit des Konzeptnetzes entsprechend, könnte man die vorkommenden Begriffe als eine Art semantischen korpusbasierten Thesaurus ansehen.

17. Wenn zwei Konzeptnetze in der bilingualen Suche gut übertragbar sind, führen sie zu den vergleichbaren relevanten Dokumenten hin. Dabei weist die Begriffsrelation darauf hin, ob es in der zugehörigen relevanten Dokumentenliste um das gleiche Thema geht.
18. Die Unsymmetrie in der Sprache erfordert manchmal etwas Sorgfalt in der Vorbereitung bei der Auswahl der Begriffe, damit die Konzeptnetze danach besser verglichen werden können.
19. Die übersetzten Zusatzbegriffe aus den relevanten Dokumenten von der Ausgangsprache sind sehr wirksam für die Suche in der Zielsprache.
20. Die Abweichung der Suchwörter führt in der Zielsprache manchmal zu nichtpassenden Dokumenten. Hier könnte man durch Beschränkung auf kleinere Textteile im Zieldokument und Analyse der Umgebung (mit der Transfermatrixtechnik) eventuell eine Verbesserung erzielen.

6.2 Ausblick

6.2.1 Sprachliche Symmetrie

Weil das Konzeptnetz die Assoziationen zwischen den Begriffen auf der Wortebene zeigt, benötigt man für die Vergleichbarkeit der Konzeptnetze zwischen den unterschiedlichen Sprachen wörtliche Isomorphe. Die Wortarterkennung, part of speech, und die Stammformreduzierung sind die Werkzeuge, die herkömmlich in dieser Situation eingesetzt werden. Durch die Wortart- und Stammformerkennung können Probleme, die von überflüssigen, wenig bedeutsamen Wörtern und von vielfältigen abgeleiteten Formen herrühren, verhindert werden. Diese Hypothese müsste aber noch einmal vertiefter überprüft werden (vgl. [ACKE00]).

Außerdem müssten die unterschiedlichen Besonderheiten wie „deutsche trennbare Verben“, „deutsche Komposita“, „englische Nominalphrase“ und „englisches Verb mit weiteren Elementen bestehender Ausdruck“ semantisch ausgeglichen werden. Einige Ansätze finden sich im Abschnitt 4.2.1. Weil die Assoziation häufig unabhängig von der Grammatik ist, kann die grammatische Struktur bei diesen Prozessen in den Hintergrund treten.

6.2.2 Grafische Darstellung

Die vom System automatisch angelegte Gruppierung der Begriffe auf dem Konzeptnetz ist für den Nutzer oft ein Zeit verbrauchendes Problem. Der Nutzer muss selbst einschätzen, welche Begriffe zu welcher Gruppe gehören, damit sich das grobe thematische Konzept einfacher gewinnen lässt. Um die Cluster besser gruppieren zu können, könnte das System die folgende Vorgehensweise verfolgen:

1. Im Clusterverfahren werden die Gruppen bestimmt. Die Anzahl der Bereiche hängt davon ab, wie viele Gruppen durch die Schwellenwertverarbeitung zusammengestellt werden.
2. Nach der Singularwertzerlegung mögen die Zentroide jeder Gruppe einen Kreismittelpunkt bilden und der Radius durch die maximale Distanz zwischen Zentroid und anderen Worten gegeben sein.
3. Die Kreise werden zur grafischen Darstellung des Konzeptnetzes verwendet.
4. Es wäre hilfreich, wenn dasselbe Thema in den verschiedenen Sprachen mit derselben Farbe dargestellt würde. Das könnte man durch eine vorherige Übersetzung und Prüfung der Zusammenhänge einiger krosslingualer Begriffe erreichen.

6.2.3 Globale Dokumentanalyse

Durch die globale Dokumentenanalyse können die Dokumente in die Gruppen aufgeteilt werden. Erst muss die Wort-Dokument-Matrix mit binären oder tf-idf Werten erzeugt werden. Hierfür hat man zwei mögliche Methoden.

1. Formale Begriffanalyse

   Diese Methode wurde von Gootjen und Van der Weide (vgl. [GRWE02] und [GRWE04] in ihrem Projekt (monolinguales Informationsretrieval) verwendet. Die Theorie stammt aus [GAWI05]. Die Relation zwischen den Indexwörtern und den Dokumenten wird dafür benutzt, um das Konzeptnetz im Sinn der formalen Begriffanalyse aufzubauen. Seine Knoten repräsentieren das 2-Tupel aus einer Menge der Begriffe und der Dokumente.

2. Cluster-Methode

   Sei $\mathbf{M} = \lfloor tf - idf_{ij} \rfloor_{nm}$, wo $i = 1,2,...,n$ die Anzahl der Indexe und $j = 1,2,...,m$ die Anzahl der Dokumente bedeutet. Sei $\mathbf{D}_{mm} = \mathbf{M}^T \cdot \mathbf{M}$. Die Matrix $\mathbf{D}$ ist die As-

soziation von den Dokumenten durch die Indexe. Die Gruppierung der Dokumente erfolgt durch eine Cluster-Methode auf der Matrix **D**.

Das Ergebnis der Gruppierung kann der SimilarDoc-Funktion oder dem Relevanz Feedback helfen. Mit der SimilarDoc-Funktion können alle Dokumente durch das Konzeptnetz ohne Echtzeitberechnung einfach gefunden werden. Beim Relevanz Feedback werden die zusätzlichen Begriffe aus dem formalen begriffsanalysierten 2-Tupel mit den Suchwörtern in die andere Sprache übertragen. Anhand der Cluster können die zusätzlichen Begriffe für das Feedback auch aufgerufen werden, indem alle in derselben Gruppe zugehörige Dokumente bzw. Spalten der Indexmatrix **M** aktiviert und die p-höchsten relevanten Begriffe durch die Berechnung $\prod_{j \in G} s_{ij}$ mit den Suchwörtern übersetzt werden, wobei *j* den Index für die Dokumente in derselben Gruppe *G* repräsentiert.

6.2.4 Relevanz-Feedback

Um das Relevanz-Feedback in die bilinguale Suche mittels Konzeptnetzen zu integrieren, müssen die geeigneten Begriffe aus den relevanten Dokumenten der Ausgangsprache herausgenommen werden und mit den Suchwörtern zusammen in die Zielsprache übersetz werden. Wenn der Suchprozess auf der Ausgangsprache beendet ist, lässt sich mit einer der beiden Methoden weiterarbeiten:

1. Nutzer-Relevanz-Feedback

   Die relevanten Dokumente aus der Trefferliste werden vom Nutzer ausgewählt.

2. Pseudo-Relevanz-Feedback

   Die ersten p-höchsten relevanten Dokumente werden automatisch vom System herausgenommen.

Danach werden r-höchste assoziierte Wörter gemäß den Suchwörtern aus den relevanten Dokumenten als zusätzliche Begriffe übernommen, mit den Suchwörtern verkettet, um das neue Konzeptnetz zu erzeugen, was dann in die Zielsprache übersetzt wird. Die r-höchsten assoziierten Wörter kommen nicht aus der globalen Analyse aller Dokumen-

te in der Sammlung, sondern aus den lokalen relevanten Dokumenten. Deswegen muss die neue Matrix für die Begriffe und ihre entsprechenden Dokumente gebildet werde. Diese Matrix kann eine einfache binäre Matrix sein. Ihre Spalten repräsentieren die Dokumente in der Sammlung und ihre Zeilen alle Indexwörter, wobei der Wert „0“ keine Beziehung zwischen dem Wort und dem Dokument und der Wert „1“ eine bestehende Beziehung markiert.

Ablauf: Es sind die r-höchsten assoziierten zusätzlichen Begriffe der relevanten Dokumente zu finden.

Voraussetzung

Sei $\mathbf{A} = \lfloor a_{ij} \rfloor_{nm}$ *die Matrix für n Indexwörter und m Dokumente.* $a_{ij} = 0$ *, wenn das Wort i keinen Zusammenhang mit dem Dokument j, und* $a_{ij} = 1$*, wenn das Wort i einen Zusammenhang mit dem Dokument j hat. Sei D die Menge der relevanten Dokumente. Sei* $\mathbf{S} \subseteq$ *gesamten indirekten Assoziationsmatrix, also* $\mathbf{S}$ *eine kleinere indirekte Assoziationsmatrix, entsprechend der Suchwörter.*

1. *Die Spaltenvektoren* $\vec{\mathbf{d}}_j$ *entsprechend der Menge D werden aktiviert. Der Anwesenheitsvektor* $\vec{\mathbf{v}}$ *wird durch* $\vec{\mathbf{v}} = \bigwedge_{j \in D} \vec{\mathbf{d}}_j$ *definiert.*

2. *Die Berechnung* $\mathbf{S} \cdot \vec{\mathbf{v}}$ *wird durchgeführt. Das Ergebnis ist der Spaltenvektor* $\vec{w} = \left[\sum_{j=1}^{n} s_{ij} \right]$.

3. *Für alle Mitglieder* $\hat{s}_i = \sum_{j=1}^{n} s_{ij}$ *werden die p Maxima* $\max_i(p) = \{\hat{s}_i\}$ *als zusätzliche Begriffe gewählt.*

6.2.5 Problem der Volltextsuche

Einige Dokumente behandeln viele Themen. Deswegen wird die TrefferDoc-Funktion, die mit Musterabgleichung arbeitet, so beeinflusst, dass auch unrelevante Dokumente als relevante aufgerufen werden. Um das zu vermeiden, muss der Suchbereich bzw. die bedeutsame Passage im Text definiert werden. Anhand der Aufspürung der Wörter in einem bestimmten Fenster und der Distanz zu den Suchwörtern, kann die Ähnlichkeit zwischen der Anfrage und einem Dokument feiner abgeschätzt werden.

Eine andere Möglichkeit ist die lokale Dokumentanalyse (vgl. [XUCR96]). Diese Methode nimmt die Dokumente vom Relevanz-Feedback und stellt eine lokale Betrachtung an. Das kann hier verwendet werden, um präzisere Informationen zu extrahieren oder unrelevante Dokumente zu unterdrücken. Die extrahierte Information kann zusammen mit den Suchwörtern dann eine neue Trefferliste erzeugen. In [XUCR96] wird gezeigt, dass die lokale Kontextanalyse, die eine globale Analysetechnik auf die lokalen Dokumente anwendet, eine verbesserte Leistungsfähigkeit und Vorhersagbarkeit bringt.

Abgesehen davon könnten die Begriffe auf dem Konzeptnetz einfach durch eine Indexmatrix verbunden werden. Die Indexmatrix kann von irgendeinem IR-Modell erzeugt werden (vgl. Abschnitt 2.2 oder [BEAZ99]).

6.2.6 Korpusbasiertes Semantiknetz

Für die Textsammlung können sich direkte und indirekte Assoziationen unter einer bestimmten Passage berechnen lassen. Mit der Worthäufigkeit können die vorkommenden Wörter samt ihren starken Assoziierten als eine extrahierte Kurzfassung gewählt werden. Solche Wörter in der Kurzfassung lassen sich nachher für ganze Sammlung verwenden. Diese Methode ist eine Möglichkeit, Informationen zu extrahieren. Die globale Analyse durch die SENTRAX kann dann das korpusbasierte Semantiknetz erzeugen.

6.2.7 Suche mittels des Konzeptnetzes für thailändische Sprache

Im Vergleich zu europäischen Sprachen sind die asiatischen Sprachen ganz anders. Bereits auf der Wortebene gibt es ganz viele Unterschiede. Bevor die bilinguale Suche zwischen einer europäischen Sprache und einer asiatischen Sprache durch das Konzeptnetz entstehen kann, muss man den sprachlichen Charakter genauer erforschen. Beim Thailändischen beispielweise, das für ca. 60 Millionen Menschen Muttersprache und Fremdsprache[26] ist, gibt es 21 Konsonanten mit 44 Formen, 32 Vokale mit mehreren Betonungen und Formen– 21 Formen, 5 Betonungen – 4 Formen.[27]

In einem Satz werden die Wörter ohne Leerzeichen und ohne Stoppzeichen am Ende geschrieben, und gibt es kein abgeleitetes Wort [SCI97]. Das Leerzeichen bedeutet in thailändischer Sprache meist das Satzende. Es wird auf drei Ebenen geschrieben, nämlich Überzone, Mittelzone und Unterzone.

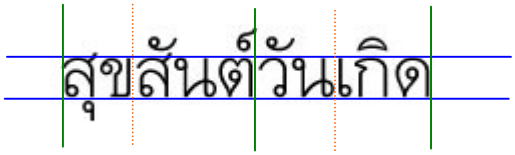


Abbildung 29 Das Beispiel für „Herzlichen Glückwunsch zum Geburtstag!“ auf thailändisch. Einige Vokale können in jeder Zone geschrieben werden. Die Buchstaben werden nur in der Mittelzone gefunden. Die durchgezogenen Vertikallinien zeigen hier die richtige Worttrennung. Die strichpunktierten Vertikallinien repräsentieren andere Möglichkeit der Trennung.

Bei einigen asiatischen Sprachen gibt es keine bestimmte Wortbegrenzung, z.B. bei der japanischen, chinesischen und thailändischen Sprache ([PSC00], [TSTC00]). Ohne gute Worttrennung ist die thailändische Sprache aber mehrdeutig. Weil die meisten thailändischen Komposita aus zwei oder mehr kleinen Tokens zusammengesetzt sind [SCI97], müssen die Worttrennung und die Wortarterkennung gut zusammenarbeiten, damit

[26] http://en.wikipedia.org/wiki/List_of_languages_by_total_speakers

[27] Von http://www.st.ac.th/bhatips/gramma3.htm

Mehrdeutigkeit durch vielfältige Trennungsmöglichkeiten verhindert wird. Die Ressource für elektronische thailändische Sprachverarbeitung findet sich in [SPWM00]. Nach einer entsprechenden sprachlichen Vorverarbeitung könnte die bilinguale Suche mittels der Konzeptnetz Technologie vermutlich auch für andere Sprachen als die hier untersuchten funktionieren.

7 ANHANG

7.1 Der TreeTagger

7.1.1 Arten des Taggeraufrufs

Mann kann den TreeTagger auf zwei verschiedene Arten, per Batchdatei oder direkt durch Befehlsaufruf, starten.

- Die Batchdatei, mit der der TreeTagger angesteuert wird, hat z.B. für Deutsch folgenden Quelletext:

```
@echo off
set TAGDIR=D:\Work\TreeTagger
set BIN=%TAGDIR%\bin
set CMD=%TAGDIR%\cmd
set LIB=%TAGDIR%\lib
set TAGOPT=%LIB%\german.par -token -lemma -sgml -no-unknown[28]
if "%2"=="" goto label1
perl %CMD%\tok-german.pl -f %LIB%\german-abbreviations %1 | %BIN%\TreeTagger
%TAGOPT% > %2
goto end
:label1
if "%1"=="" goto label2
perl %CMD%\tok-german.pl -f %LIB%\german-abbreviations %1 | %BIN%\TreeTagger
%TAGOPT%
goto end
:label2
echo.
echo Usage: tag-german file {file}
echo.
:end
```

- Der direkte Befehlsaufruf wird in diesem Sinne verwendet:

  tree-tagger {-options-} <parameter file> {<input file> {<output file>}}

[28] Diese Optionen werden in dieser Arbeit verwendet. Das Ausgabeformat des Taggers ist dadurch dreispaltig (Wort | Tag | Stamm).

7.1.2 Argumente

- **parameter file:** Der Name der Parameterdatei (englisch.par oder german.par).
- **input file:** Der komplette Pfad der Eingabedatei. Diese Datei wird durch die Perl-Anwendung so umgewandelt, dass nur ein Wort pro Zeile steht. Die Token-Übersetzungsdatei, tok-english.pl und tok-german.pl, erzeugt dann das richtige Eingabeformat.
- **output file:** Der komplette Pfad der Ausgabedatei. Die Ausgabe liefert ein Wort pro Zeile. Die Ausgabezeilen können dabei durch Tabulatoren (Tab) getrennt mehrspaltig sein. In diesen Spalten können z.B. neben dem Wort auch das zugehörige Tag und die Stammform stehen.

7.1.3 Optionen

Verwendete Optionen

- -**token**: Originalwort in der ersten Spalte der Ausgabe angeben.
- -**lemma**: Stammform angeben /erzeugen.
- -**sgml**: tag SGML nicht angeben, bsw. Anfang der Zeile mit '<' und Ende mit '>'.
- -**no-unknown:** wenn kein Stammwort ermittelbar ist, dann das Originalwort in der Stammwortspalte der Ausgabe angeben.

Einige Beispiele für weitere mögliche Optionen

- -**threshold <p>**: Ausgabe nur dann, wenn das zugehörige Tag eine höhere Wahrscheinlichkeit als <p> hat.
- -**prob**: Tagwahrscheinlichkeit mit ausgeben. (erfordert die Option -threshold)
- -**no-heuristics:** keine Heuristik auf dem Lexikon anwenden.
- -**quiet**: keine Statusnachrichten ausgeben.

Außer den obengenanten Optionen gibt es noch weitere, die hier nicht erwähnt werden.

7.1.4 Markierungen des TreeTaggers

7.1.4.1 Deutsche Markierungen im TreeTagger

POS=	Beschreibung	Beispiele
ADJA	attributives Adjektiv	[das] große [Haus]
ADJD	adverbiales oder prädikatives Adjektiv	[er fährt] schnell [er ist] schnell
ADV	Adverb	schon, bald, doch
APPR	Präposition; Zirkumposition links	in [der Stadt], ohne [mich]
APPRART	Präposition mit Artikel	im [Haus], zur [Sache]
APPO	Postposition	[ihm] zufolge, [der Sache] wegen
APZR	Zirkumposition rechts	[von jetzt] an
ART	bestimmter oder unbestimmter Artikel	der, die, das, ein, eine
CARD	Kardinalzahl	zwei [Männer], [im Jahre] 1994
FM	Fremdsprachliches Material	[Er hat das mit "] A big fish ["übersetzt]
ITJ	Interjektion	mhm, ach, tja
KOUI	unterordnende Konjunktion mit \zu" und Infinitiv	um [zu leben], anstatt [zu fragen]
KOUS	unterordnende Konjunktion mit Satz	weil, daß, damit, wenn, ob
KON	nebenordnende Konjunktion	und, oder, aber
KOKOM	Vergleichspartikel, ohne Satz	als, wie
NN	normales Nomen	Tisch, Herr, [das] Reisen
NE	Eigennamen	Hans, Hamburg, HSV
PDS	substituierendes Demonstrativpronomen	dieser, jener
PDAT	attribuierendes Demonstrativpronomen	jener [Mensch]
PIS	substituierendes Indefinitpronomen	keiner, viele, man, niemand
PIAT	attribuierendes Indefinitpronomen ohne Determiner	kein [Mensch], irgendein [Glas]
PIDAT	attribuierendes Indefinitpronomen mit Determiner	[ein] wenig [Wasser], [die] beiden [Brüder]
PPER	irreflexives Personalpronomen	ich, er, ihm, mich, dir
PPOSS	substituierendes Possessivpronomen	meins, deiner
PPOSAT	attribuierendes Possessivpronomen	mein [Buch], deine [Mutter]
PRELS	Relativpronomen substituierend	[der Hund,] der
PRELAT	Relativpronomen attribuierend	[der Mann ,] dessen [Hund]
PRF	reflexives Personalpronomen	sich, einander, dich, mir
PWS	substituierendes Interrogativpronomen	wer, was
PWAT	attribuierendes Interrogativpronomen	welche [Farbe], wessen [Hut]
PWAV	adverbiales Interrogativ oder Relativpronomen	warum, wo, wann, worüber, wobei
PAV	Pronominaladverb	dafür, dabei, deswegen, trotzdem

PTKZU	"zu" vor Infinitiv	zu [gehen]
PTKNEG	Negationspartikel	nicht
PTKVZ	abgetrennter Verbzusatz	[er kommt] an, [er fährt] rad
PTKANT	Antwortpartikel	ja, nein, danke, bitte
PTKA	Partikel bei Adjektiv oder Adverb	am [schönsten], zu [schnell]
TRUNC	Kompositions-Erstglied	An- [und Abreise]
VVFIN	finites Verb, voll	[du] gehst, [wir] kommen [an]
VVIMP	Imperativ, voll	komm [!]
VVINF	Infinitiv, voll	gehen, ankommen
VVIZU	Infinitiv mit "zu", voll	anzukommen, loszulassen
VVPP	Partizip Perfekt, voll	gegangen, angekommen
VAFIN	finites Verb, aux	[du] bist, [wir] werden
VAIMP	Imperativ, aux	sei [ruhig !]
VAINF	Infinitiv, aux	werden, sein
VAPP	Partizip Perfekt, aux	gewesen
VMFIN	finites Verb, modal	dürfen
VMINF	Infinitiv, modal	wollen
VMPP	Partizip Perfekt, modal	[er hat] gekonnt
XY	Nichtwort, Sonderzeichen enthaltend	D2XW3
$,	Komma	,
$.	Satzbeendende Interpunktion	. ? ! ; :
$(	sonstige Satzzeichen; satzintern	- [ ]()

7.1.4.2 Englische Markierungen im TreeTagger

POS =	Beschreibung	Beispiele
CC	Coordinating conjunction	and, but, nor, or, yet, plus, minus, less, times
CD	Cardinal number	one
DT	Determiner	a(n), any, another, some, each
EX	Existential there	There/EX was a party inprogress
FW	Foreign word	persona non grata
IN	Preposition or subordinating conjunction	because/IN of/IN her late arrival
JJ	Adjective	one-of-a-kind, fourth, full
JJR	Adjective, comparative	larger, more
JJS	Adjective, superlative	smallest, most, least
LS	List item marker	Leters and numberals which are used to identify items in a list.
MD	Modal	should, can, may
NN	Noun, singular or mass	income/NN tax/NN return, that´s a nice red/NN, Good cooking/NN is something to enjoy
NNS	Noun, plural	The police/NNS have arrived on the scene

NP	Proper noun, singular	John/NP ´s/POS idea
NPS	Proper noun, plural	the parents/NNS
PDT	Predeterminer	All/PDT his marbles
POS	Possessive ending	John/NP ´s/POS idea
PP	Personal pronoun	I, me, you, he, -self or –selves, mine, yours, his, her
PP$	Possessive pronoun	My, your, her, its
RB	Adverb	one-half/RB the amount, They won hardily/RB
RBR	Adverb, comparative	I can´t run any/RB further/RBR, We are closer/RBR to home.
RBS	Adverb, superlative	most every-
RP	Particle	She told off/RP her friends.
SYM	Symbol	mathematical, scientific and technical symbols.
TO	to	to
UH	Interjection	oh, please,well
VB	Verb, base form	do
VBD	Verb, past tense	were
VBG	Verb, gerund or present participle	Concerning/VBG your request of last week.
VBN	Verb, past participle	Provided/VBN that he comes.
VBP	Verb, non-3rd person singular present	come, take, run
VBZ	Verb, 3rd person singular present	looks, makes
WDT	Wh-determiner	A man that/WDT I know
WP	Wh-pronoun	Tell me what/WP you would like to eat.
WP$	Possessive wh-pronoun	whose
WRB	Wh-adverb	However/WRB much he wants to, he can´t.

7.2 SENTRAX-Engine

7.2.1 Die Funktionen der SENTRAX

7.2.1.1 LexicoMap

Diese Funktion begegnet dem Problem der Schreibweisenvarianten und Tippfehler. Herkömmliche IR-Systeme führen ein Matching des Eingabestrings mit den Einträgen in der „invertierten“ Wortliste aus. Dabei kann es zu „mismatches“ kommen, die auf einer Vielzahl von Gründen beruhen. Zum Beispiel kann es Tippfehler in der Eingabe geben, auch Tippfehler im Text. Oder es gibt zulässige (oder einfach oft benutzte) Schreibvarianten, wie Potenzial - Potential oder Appartement – Appartment – Apartement – Apartment oder fremdsprachige Namen, wie z.B. Tschebyscheff-Chebychev. Auch kann es unterschiedliche Beschreibungen geben, obwohl dasselbe gemeint ist, wie z.B. Uranbergbau-Uranerzbergbau.

Die LexicoMap bietet hier Abhilfe durch eine fehler- und Varianten-tolerante Suche basierend auf Stringähnlichkeit. Sie verzeiht also Tippfehler, OCR-Fehler oder Schreibvarianten, wie sie in gewöhnlichen Suchanfragen häufig vorkommen. Die LexicoMap ist auch imstande Kompositabildungen und ihre Variationen zu finden, sobald die Stammform in der Eingabe ist – und umgekehrt.

Die folgende Abbildung zeigt das Beispiel „Nahostkonflikt“ in einer fehlerhaften Variation. Die LexicoMap zeigt ähnliche Begriffe sowie Komposita.

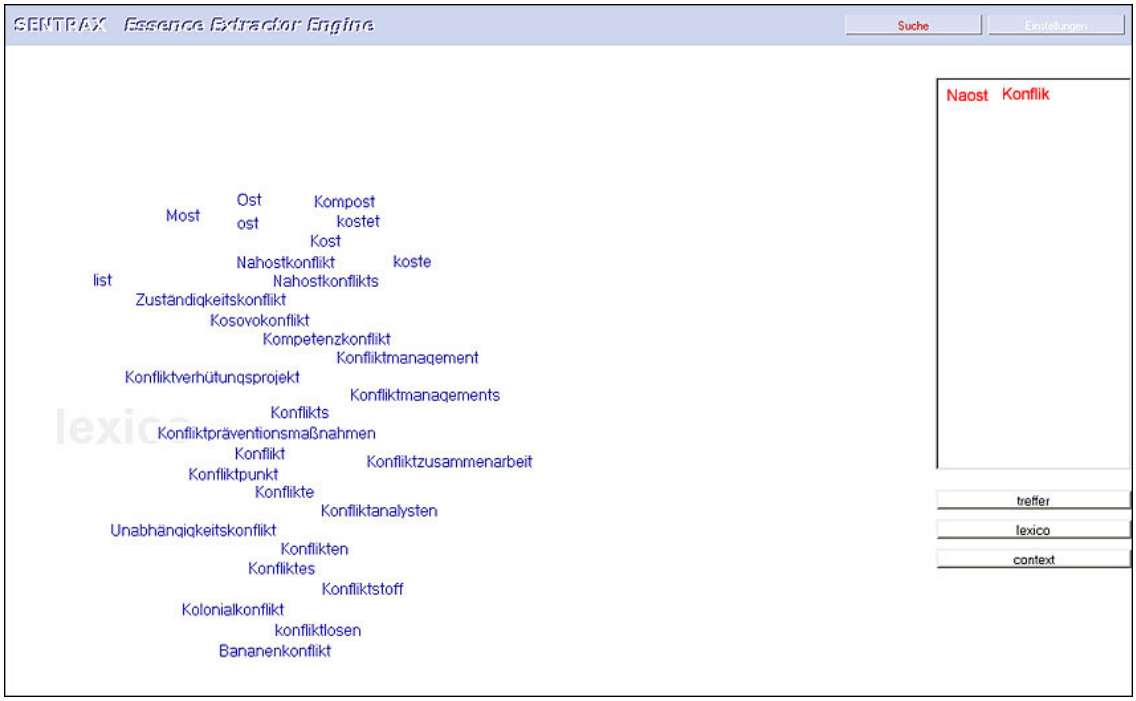


Abbildung 30. LexicoMap: Eingabe: Naost und Konflik (jeweils Tippfehler).

Abbildung 31 Andere Ansichtoption der LexicoMap als Liste.

7.2.1.2 ContextMap

Die ContextMap ermöglicht eine semantische oder Konzept-orientierte Suche, indem sie in der Datensammlung häufig gemeinsam auftretende Begriffe statistisch analysiert. Der Datenbestand wird vollautomatisch auf Wortkookkurrenzen untersucht, das Resultat als ContextMap-Index abgespeichert. So können direkte und indirekte Assoziationen innerhalb des gesamten Korpus als Sinnstrukturen erschlossen und in Form von „Begriffswolken“ dargestellt werden. Im Gegensatz zu übergreifenden Methoden (wie z.B. der Verwendung von Thesauri) enthält die so entstehende Darstellung stets nur Wörter aus dem Korpus. Der Vorteil dieser Beschränkung ist, dass sich so auch ein grober Zusammenhang zwischen den Dokumenten der Datenbasis erkennen lässt. Dies unterstützt den Suchenden nicht nur beim Finden erweiternder Suchbegriffe, sondern fördert auch Verständnis für den Korpus.

Abbildung 32 zeigt das Beispiel „Nahost“ und darum gruppierte assoziierte Begriffe. Man kann sich damit ein Bild machen, welche Themen die betroffenen Dokumente beschreiben.

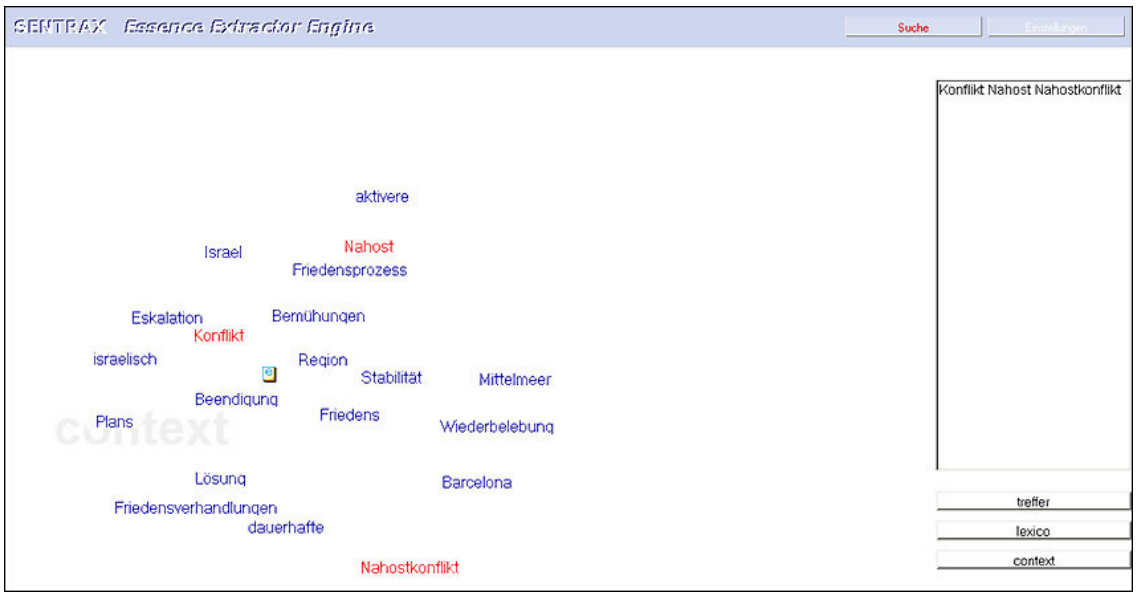


Abbildung 32 ContextMap nach der Auswahl der Attribute „Konflikt, Nahost, Nahostkonflikt“.

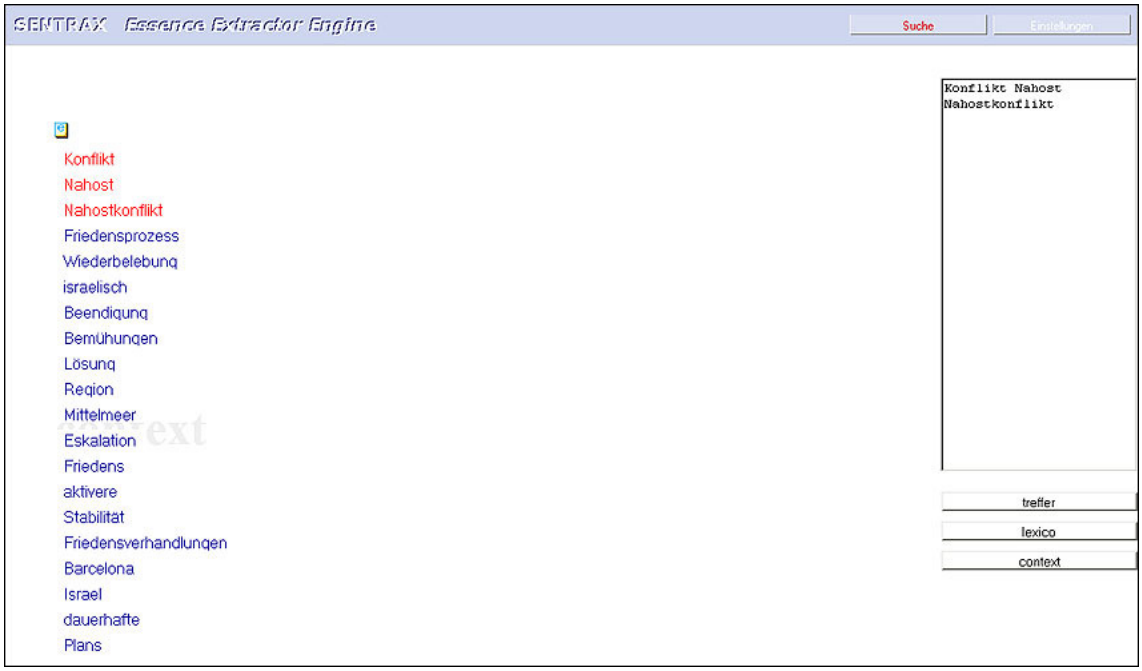


Abbildung 33 ContextListe ist eine Ansichtoption der ContextMap, deren Attribute nach ihrer Assoziationsstärke sortiert werden.

7.2.1.3 TrefferDoc und Ansichtsoptionen eines Dokuments

Die Option TrefferDoc liefert Dokumente entsprechend den Suchwörtern als Liste zurück. Dies entspricht einer herkömmlichen Trefferliste. Dabei wird standardmäßig von allen Dateien, über die ein Index gebildet wird, eine Kopie im HTML-Format erzeugt und auf der Festplatte gespeichert. Das ermöglicht 3 Ansichtoptionen eines Dokuments:

1 Dokument im Originalformat (unterstrichener Link)

2 Dokument im HTML-Format mit Highlight-Funktion (Symbol)

3 Das Dokument im HTML-Format mit Click-Highlight-Funktion ohne Bilder und ohne Hyperlinks. Hier verweist jedes Suchwort auf das nächstfolgende Suchwort im Dokument; man gelangt per Mausklick auf ein Suchwort automatisch zur Fundstelle des nächsten Suchworts. (Symbol).

Die 3 Ansichtoptionen findet man in der Trefferliste:

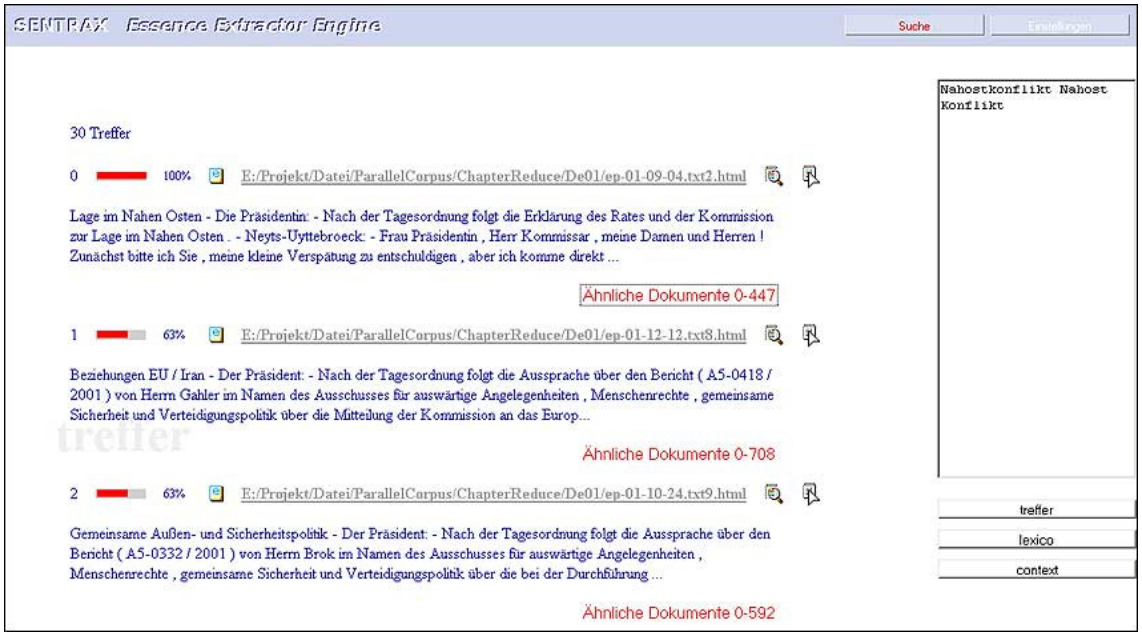


Abbildung 34 TrefferListe wird durch die Treffer-Funktion erzeugt.

7.2.1.4 SimilarDoc

Diese Funktion ermöglicht die Suche nach einander ähnlichen Dokumenten im Bestand. SimilarDoc-Funktion kann erst nach Erhalt einer Trefferliste aktiviert werden. Der Nutzer wählt die bei dem (beliebigen) Trefferdokument mitgeführte Option „Ähnliche Dokumente NN““ und erhält eine neue Trefferliste, diesesmal sortiert nach Dokumentenähnlichkeit.

Abbildung 35 Nach dem Mausklick auf „Ähnlichkeit Dokument“ wird die SimilarDoc-Funktion aktiviert. Die Ähnlichkeit zwischen dem gewählten Dokument und den anderen Dokumenten wird berechnet.

7.2.2 Die Ähnlichkeitsmaße

Während die LexicoMap-Funktion String-orientiert arbeitet und hauptsächlich auf der Basis von n-Grammen arbeitet, findet die ContextMap bedeutungsverwandte Begriffe in den Dokumenten. Dies beruht auf der Auswertung von Auftretenshäufigkeiten und nahem Beieinanderstehen von Worten und Wortgruppen in den Texten. Man hat daher oft semantisch verwandte Begriffe in der ContextMap, wie z.B. *Fusion-Zusammenschluss*, es werden aber auch gänzlich verschiedene Worte dort zusammengebracht, wie z.B. *Ausbildung-Analphabetentum*, weil sie durch die Art ihres Auftretens in den Dokumenten einen Vorgang oder eine Idee repräsentieren. Die Güte dieser Funktion hängt von der Homogenität des Datenmaterials ab. Für normale Texte, die aus ordentlichen Sätzen bestehen, funktioniert die ContextMap ziemlich gut. Für Wörter, die inhaltlich zusammenhangslos in Tabellen stehen, wie z.B. in Telefonlisten, darf nicht zuviel von der ContextMap-Funktion erwartet werden, da der „Kontext“ vom Benutzer nicht zuverlässig interpretiert werden kann.

Die TrefferDoc-Funktion zeigt alle Dokumente, in denen die Suchwörter enthalten sind mit 100% an. Im Falle des Fehlens einiger Eingabebegriffe wird die Ausgabeliste entsprechend modifiziert, so dass ein Dokument mit solchen Mängeln eine Rangabstufung erfährt. Dokumente auf gleicher Stufe werden nach ihrer intern vergebenen ID sortiert. Innerhalb einer festen Prozentgruppe sind also alle Dokumente gleich gut.

Die SimilarDoc-Funktion arbeitet wieder auf den Wörtern (jetzt des gesamten Textes) und sucht entsprechend passende Dokumente zusammen. Auch hier sind alle Treffer auf derselben Prozentstufe gleichermassen gut. Diese Funktion ist nicht notwendig symmetrisch, was aber das Empfinden des Benutzers eigentlich nicht stören sollte. Denn auch ohne IR-Systeme kann es vorkommen, dass ein Dokument A bestpassend zum Dokument B ist, B wiederum (weil es vielleicht viel umfangreicher als A ist) besser zu C passt.

7.3 TIHO-Anwendung

Die Hilfssoftware TIHO (Abk. für „**T**agged**I**n**H**tml**O**ut“) wird parallel zu dieser Arbeit im Rahmen einer Diplomarbeit im Bereich Informatik mit dem Titel „Automatisierte Wortlistenerzeugung durch multiple Dokumentenreduzierung im bilingualen Kontext“ entwickelt. Daher wird hier die Version von TIHO beschrieben, die im Rahmen der Vorbereitung der Container für die Untersuchungen in dieser Arbeit verwendet wurde.

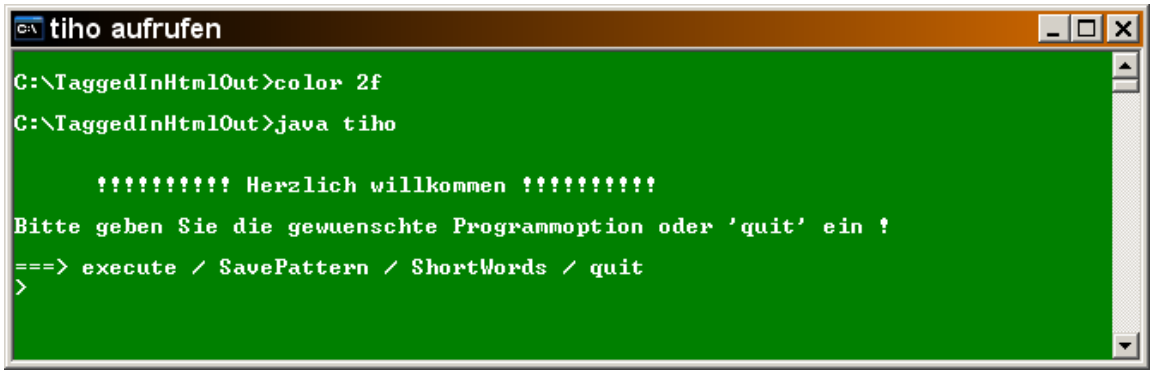
Die Benutzeroberfläche ist im Moment auf MS-DOS-Ebene angelegt, da es bei der Entwicklung von TIHO zunächst um Funktionalität und dann erst um Komfort geht.

TIHO wurde zunächst speziell für die Einstellungen bzw. Optionen des TreeTaggers entwickelt, die in dieser Arbeit Verwendung finden, da die vom TreeTagger erzeugten Dateien in ihrer Spaltenanzahl und Spaltendarstellung je nach Option variieren (siehe Abschnitt 7.1.1).

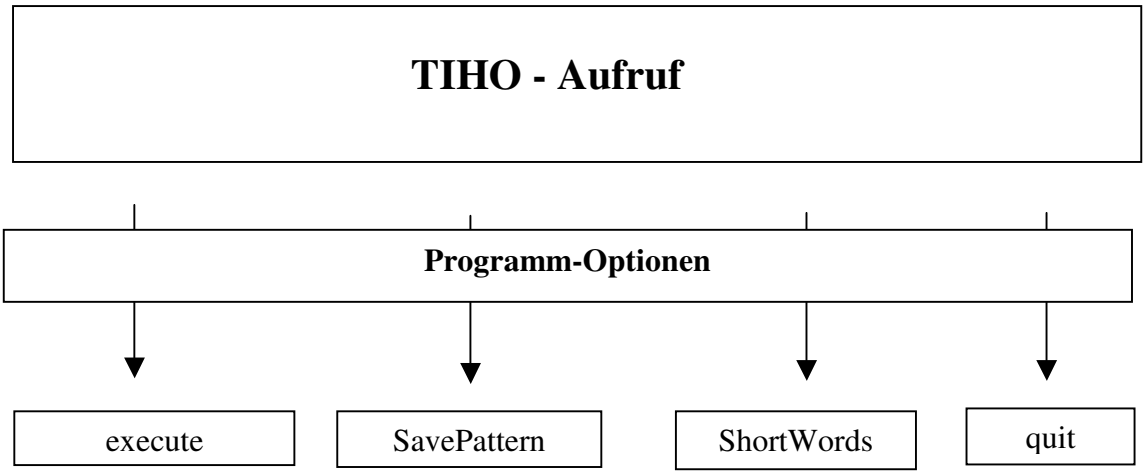
TIHO wird mit Hilfe folgender Batchdatei aufgerufen, die in dem TIHO-Programmordner liegt:

color 2f : Hintergrundfarbe grün für “TIHO wurde erfolgreich aufgerufen”
java tiho : Aufruf von tiho.class mit dem java-Interpreter
color 4f : Hintergrundfarbe rot für “TIHO wurde beendet”
pause : Um die Meldungen lesen zu können. Beendung erst bei Drücken einer beliebigen Taste.

Nach dem Aufruf erscheint eine zu dieser folgenden analoge Anzeige auf dem Bildschirm:



TIHO umfasst somit folgende Funktionen:



7.3.1 Beschreibung der Funktionen von TIHO

7.3.1.1 Execute

Durchlauf auf einem beliebigen Quellverzeichnis inkl. Unterstruktur mit:

- Automatisiertem Taggen
  Wahlweise mit oder ohne Verzeichnisstrukturerzeugung und mit oder ohne Erzeugung einzelner tag-Dateien für jede gefundene Dokumentdatei im Programmunterordner „tagFiles“ analog zur Quellstruktur.

- Anwendung von „Löschoptionen“
  - Alle Worte bis zu einer wählbaren Länge generell aus dem Text entfernen.
  - Anwendung (nicht zwingend) eines (tag-) Schemas wahlweise im Sinne „lösche“ oder „behalte“ im Text.
  - Anwendung (nicht zwingend) einer „Löschwortliste“ / Stoppwortliste

- Erzeugung der „Restdokumente“ (inkl. Verzeichnisstruktur analog zur Quelle)
  (Sonderfall: Quelle txt-File → Ergebnis html-File !)
  - 1x mit den Worten (in „-W“).
  - 1x mit den tags (in „-T“).
  - 1x mit den Stammworten (in „-SW“).

  Erzeugung im Ziel mit Quellordnername + „-W“ / + „-T“ / + „-SW“ als neue ROOT-Namen.

  Falls vom Benutzer gewählt, zusätzlich: (Erzeugung im TIHO-Programmunterordner „RestHäufigkeitslisten“)

- Erzeugung von „RestWortListe / -n“ in 4-Spalten-Darstellung
  i) Wort im Text
  ii) tag
  iii) Stamm
  iv) Häufigkeit (= Anzahl Vorkommen, nicht in % !)

  wählbare Optionen hierfür :
  - eine Liste über alle Dokumente
  - eine Liste pro Dokument (inkl. Erzeugung der Unterstruktur am Zielort)

- Erzeugung von „RestStammWortListen“ in 2-Spalten-Darstellung
  i) Stamm
  ii) Häufigkeit / Anzahl

  wählbare Optionen hierfür :
  - eine Liste über alle Dokumente
  - eine Liste pro Dokument (inkl. Erzeugung der Unterstruktur am Zielort)

7.3.1.2 SavePattern

Erzeugung eines (tag -) Schemas in Form einer HTML – Seite im Programmunterordner „Muster“.
Beispiel.:
<html><body>NN
NNS
NP
NPS
</body></html>
ist der HTML-Quelltext für ein Schema, das die englischen tags “NN”, “NNS”, “NP”, “NPS” beinhaltet. Dieses Schema ist dann später in der Option „execute“ als Lösch- oder Bleibschema verwendbar.

7.3.1.3 ShortWords

Erzeugung von Listen der Worte mit den Längen 1,2 & 3 im Programmunterordner „KurzWortListen“ über alle Dokumente in der Quellstruktur mit:

- Man kann wählen ob die Listen für die Wortlängen 1 oder 1, 2 oder 1, 2, 3 angelegt werden sollen.
- Ein Schema aus „Muster“ kann angewendet werden, um nur die Worte in der Liste aufzunehmen, die nicht im Schema berücksichtigt wurden, da die anderen schon durch das Schema in „execute“ gelöscht werden könnten.

Diese Option soll dabei helfen später (im Moment) manuell (z.B. mit EXCEL) Stoppwortlisten zu erzeugen, um diese dann in „execute“ anzuwenden. Die Idee, die dahinter steht ist die, dass „Wörter“ mit bis zu 3 Zeichen nicht oder nur sehr schwer bzgl. ihrer Rechtschreibung zu beurteilen sind (Beispiel: CDU ↔ CSU).

An dieser Stelle sei kurz beschrieben, wie man zwecks Stoppwortlistenerzeugung fortfahren müsste:

- mit „ShortWords“ erzeugte Liste in EXCEL laden
- dort entsprechend verkürzen
- abspeichern als „Text (Tabstopp-getrennt)“ im TIHO-Unterordner „StoppWortListen“.

Anzumerken ist dabei, das einige Zeichenkombinationen in EXCEL andere Bedeutung zu haben scheinen.

Zum Beispiel kann aus

wort	tag	stamm	Anzahl
"	$(	"	151

als „Tabstopp-getrennt“- Datei

wort	tag	stamm	Anzahl
""""	$(	""""	151

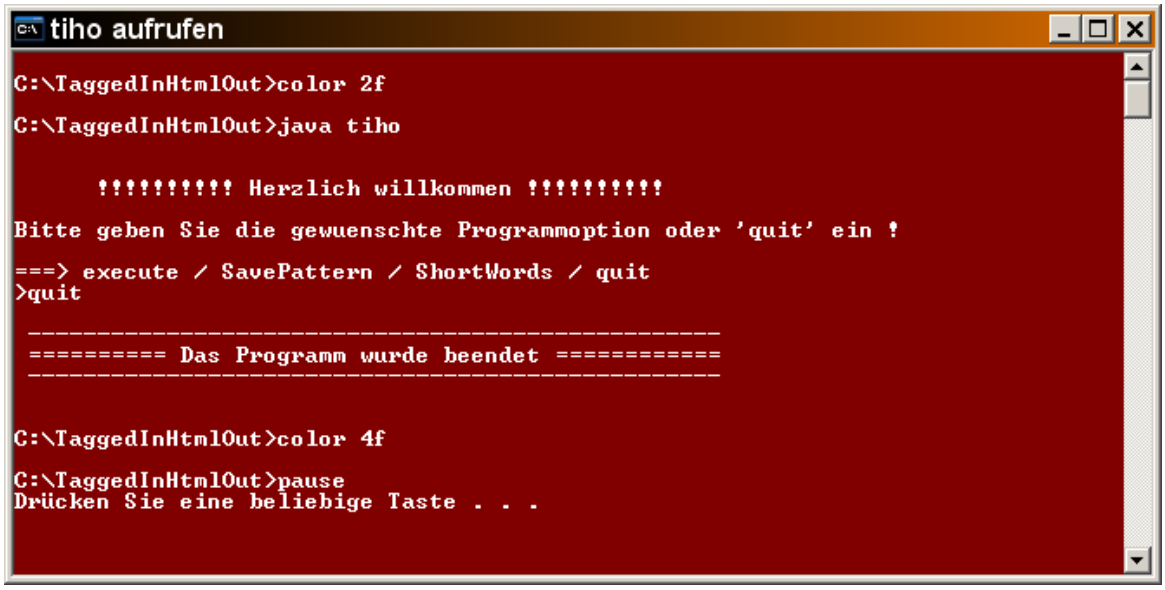
entstehen.

7.3.1.4 quit

Eine Programm – Beendung mit „quit“ ist aus Sicherheitsgründen direkt oder in fast jedem Schritt der anderen Optionen möglich.

7.3.2 Beendung von TIHO

Wurde TIHO beendet, dann erscheint eine Ausgabe analog zu dieser:



Um TIHO endgültig zu beenden ist nun nur noch eine beliebige Taste zu drücken.

7.3.3 Angestrebte Erweiterung von TIHO

- Dateinamen, die in MS-DOS teilweise nicht zugelassen sind oder umcodiert werden müssen (z.B. mit Leerzeichen, Umlauten … etc.), werden im Moment noch nicht herausgefiltert bzw. entsprechend behandelt bei der Bildung der Batch-Datei zwecks TreeTagger – Aufruf in „execute“ und „Shortwords“.
- Ein bilingualer Dokumenten-Container ist im Moment nicht bearbeitbar, da noch kein Algorithmus zur Spracherkennung eingebaut wurde. Der Benutzer muss bei der hier beschriebenen Version bei den Dialogen in den Optionen „execute“ und „ShortWords“ angeben, ob der TreeTagger für deutsche oder für englische Texte aufgerufen werden soll.
- Möglichkeit zum Abgleich der Wortlisten mit dem Wörterbuch
- komfortable Benutzeroberfläche
- Eine Möglichkeit Einstellungen (z.B. Programmpfad, Quellpfad, Mustername, … etc.) abzuspeichern (z.B. als ini-Datei) und zu laden um die Anzahl der notwendigen Nutzereingaben zu minimieren.

7.4 Sonstige Tabelle

7.4.1 Liste der englischen Nomenpräposition

A	access to, advantage of, admiration for, alternative to, attack on, attitude to /towards, authority on, association sth. with sth.
C	commend on, comparison between, connection between, contrast with, credit for, cruelty towards, charasteristic of, cure for
D	decrease in, delay in, desire for, difference between/of, difficulty in/with, disadvantage of
E	effect on, exception to, expert on/at/in, experience in
H	hope for
I	increase in, influence on, information about, intention of
K	Knowledge
L	lack of, link with
M	matter with
N	need for, notice of
O	opinion of/about
P	pleasure in, preference for, protection from
R	reaction to, reason for, recipe for, reduction in, relationship with, report on, responsibility for, result of, respect for, rise in, room for
S	solution to, smell of, sympathy for
T	tax on, taste of, threat to, trouble with
U	use of
V	victims of

7.4.2 Liste des englischen Phrasal-Verbes

A	act up, act like, add up (2), add up to, ask out
B	back down, back off, back up (4), beg off, blow up (3), bone up on, break down (2), break in(to) (3), break up (2), bring (take) back, bring off, bring up (2), brush up on, build up, burn down, burn up (2), butt in, butter up
C	call off, call on, calm down, (not) care for, care for, catch on, catch up (with), check in(to), check off, check out (of), check out, cheer up, chew out, chicken out, chip in, clam up, come across, come down with, come to (2), count on, crack down (on), cross out, cut back (on)
D	do in, do over, drag on, draw up, drop off, drop in (on), drop by, drop out (of), draw out
E	eat out , egg on, end up (2)

F	face up to, fall through, feel up to, figure out, fill in (2), fill in for, fill out (2), find out (about)
G	get across, get along (with), get around (2), get around to, get by, get in (2), get on, get off (3), get out of (2), get over (2), get rid of, get up, give up (2), go out with, go with (2), goof off, grow up
H	hand in, hand out, hang up, have to do with, hold up (3)
I	iron out
J	jack up (2), jump all over
K	keep on (2), kick out, knock out, knock oneself out
L	lay off, leave out, let down, let up, look back on, look down on, look forward to, look in on, look into, look like, look over, look up (2), look up to, luck out
M	make fun of, make up (2), make up (with), make out, make for (2), mark up, mark down, mix up
N	nod off
P	pan out, pass away, pass out (2), pick out, pick up (4), pick on, pitch in, pull off, pull over, put away, put off, put on (2), put out, put up (2), put up with, put back
R	rip off, round off, run into, run out of
S	set up, set back (2), slip up, stand out, stand up (2), show up (2), stand for (2)
T	take after, take / bring back, take care of (2), take off (3), take up, tell someone off, tick off (2), throw away, throw out (2), throw up, try on, try out, try out (for), turn around (3), turn in (3), turn down (2), turn off (2), turn on (2), turn up (2)
W	wait on (2), wake up (2), watch out for, wear out, work out (2), wrap up (3), write up, write down
Z	zonk out

Von http://www.eslcafe.com/pv/pv-list.html

(n) bedeutet, dass das Phrasal-Verb n verschiedene Bedeutungen haben kann

7.5 Formeln

Formel 1

$$w_{i,j} = f_{i,j} \times idf_i$$

Formel 2

$$f_{i,j} = \frac{freq_{i,j}}{\max_l freq_{l,j}},$$

wobei $freq_{i,j}$ die Häufigkeit des existenten Indexes *i* im Dokument *j* und

$\max_l freq_{l,j}$ das Maximum der Häufigkeit aller existenten Indexierungen im Dokument *j* ist.

Formel 3

$$idf_i = \log \frac{N}{n_i}$$

wobei N die Anzahl der Dokumente in der Sammlung ist und

n_i die Anzahl der Dokumente ist, in denen die Indexierung *i* existiert.

Formel 4

$$sim(d_j, q) = \frac{\vec{d_j} \bullet \vec{q}}{\left|\vec{d_j}\right| \times \left|\vec{q}\right|} \quad \text{oder} \quad sim(d_j, q) = \frac{\sum_i w_{i,j} \times w_{i,q}}{\sqrt{\sum_i w_{i,j}^2} \times \sqrt{\sum_i w_{i,q}^2}}$$

Formel 5

$$w_{i,q} = \left(0{,}5 + \frac{0{,}5 \cdot freq_{i,q}}{\max_l freq_{l,q}}\right) \times \log \frac{N}{n_i},$$

wobei $freq_{i,q}$ die Häufigkeit des Indexterms i in der Anfrage q ist.

Formel 6

$$P(k_i \mid R) = 0{,}5$$

$$P(k_i \mid \overline{R}) = \frac{n_i}{N},$$

wobei $P(k_i \mid R)$ die Wahrscheinlichkeit des getroffenen Indexes k_i in einem zufällig ausgewählten Dokument in R repräsentiert und

$P(k_i \mid \overline{R})$ die Wahrscheinlichkeit des getroffenen Indexes k_i in einem zufällig ausgewählten Dokument in $\overline{R}$ repräsentiert.

n_i die Anzahl der Dokumente, in den sich der Index k_i befindet.

N die Anzahl der Dokumente im Korpus.

Formel 7

$$P(k_i \mid R) = \frac{V_i + \frac{n_i}{N}}{V + 1}$$

Formel 8

$$P(k_i \mid \overline{R}) = \frac{n_i - V_i + \frac{n_i}{N}}{N - V + 1}$$

Formel 9

$$sim(d_j, q) \sim \sum_i w_{i,q} \times w_{i,j} \times \left(\log \frac{P(k_i \mid R)}{1 - P(k_i \mid R)} + \log \frac{1 - P(k_i \mid \overline{R})}{P(k_i \mid \overline{R})} \right)$$

Formel 10

$$w_{x,a} = f_{x,a} \frac{idf_x}{\max_i idf_i}$$

wobei $f_{x,a}$ die normalisierte Häufigkeit der Indexierungsausdrücke k_x im Dokument a repräsentiert,

idf_x die invertierte Dokumentshäufigkeit für die Indexierungsausdrücke k_x repräsentiert,

$\max_i idf_i$ das Maximum von $idf_i, \forall i = 1,\ldots,t$ repräsentiert.

Formel 11

$$Sim(q_{or}, d) = \sqrt{\frac{x^2 + y^2}{2}}$$

$$Sim(q_{and}, d) = 1 - \sqrt{\frac{(1-x)^2 + (1-y)^2}{2}}$$

Formel 12

$$Sim(q_{or}, d) = \left(\frac{\sum_{i=1}^{m} x_i{}^p}{m} \right)^{\frac{1}{p}}$$

$$Sim(q_{and}, d) = 1 - \left(\frac{\sum_{i=1}^{m} (1 - x_i)^p}{m} \right)^{\frac{1}{p}}$$

wobei *m* die Anzahl der in der disjunktiven oder konjunktiven Anfrage gefundenen Stichwörter ist.

Formel 13

$$\overrightarrow{k_i} = \frac{\sum_{\forall r, g_i(m_r)=1} c_{i,r} \overrightarrow{m_r}}{\sqrt{\sum_{\forall r, g_i(m_r)=1} c_{i,r}^2}}$$

$$c_{i,r} = \sum_{d_j | g_l(\overrightarrow{d_j}) = g_l(m_r) \text{ for all } l} w_{i,j}$$

wobei $g_i\left(\vec{x}\right) = 1$, wenn die Komponente i auf dem Platz i des Vektors $\vec{x}$ „1“ ist, sonst „0“.

Formel 14

$$\mathbf{M} = \mathbf{KSD}^t$$

wobei $\mathbf{K}$ die Matrix des Eigenvektors ist, der aus der Korrelationsmatrix der Ausdrücke $\mathbf{MM}^t$ hergeleitet wird,

$\mathbf{D}^t$ die transponierte Matrix des Eigenvektors ist, der aus der Korrelationsmatrix der Dokumente $\mathbf{M}^t\mathbf{M}$ hergeleitet wird,

$\mathbf{S}$ die diagonale $r \times r$ Matrix des Singularwert ist mit $r = \min(t, N)$ als Rang der Matrix $\mathbf{M}$.

Formel 15

$$\mathbf{M}_s = \mathbf{K}_s \mathbf{S}_s \mathbf{D}_s^t$$

Formel 16

$$Avg.\ Prec = \sum_{i=1}^{|q|} \frac{Prec_{q_i}\left(Recall\right)}{|q|}$$

Formel 17

$$Prec(recall_j) = \max_{recall_j \le recall \le recall_{j+1}} Prec\left(recall\right)$$

Formel 18

$$E(j) = 1 - \frac{1 + b^2}{\frac{b^2}{recall_j} + \frac{1}{prec_j}}$$

wobei $E(j)$ das E-Maß bezüglich $recall_j$ und $prec_j$ ist.

$recall_j$ der Recallwert an der Stelle *j* in der Rangliste ist.

$prec_j$ der Precisionwert an der Stelle *j* in der Rangliste ist.

b der Parameter ist, der vom Nutzer abhängig ist.

Dabei gilt $b > 1$, wenn die Precision wichtiger als der Recall ist,

$b = 1$, wenn die Precision genau so wichtig wie der Recall ist,

$b < 1$, wenn die Precision weniger wichtig als der Recall ist.

Formel 19

$$\vec{q}_{ideal} = \frac{1}{|C_r|} \sum_{\forall d_j \in C_r} \vec{d}_j - \frac{1}{N - |C_r|} \sum_{\forall d_j \notin C_r} \vec{d}_j$$

wobei C_r die Menge der relevanten Dokumente in der Sammlung und

$|C_r|$ die Anzahl der relevanten Dokumente in der Sammlung repräsentiert.

Formel 20

$$\vec{q} = a\vec{q} + b \sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - c \sum_{\forall \vec{d}_j \in D_n} \vec{d}_j$$

Formel 21

$$\vec{q}_{Rochio} = \alpha\vec{q} + \frac{\beta}{|D_r|} \sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \frac{\gamma}{|D_n|} \sum_{\forall \vec{d}_j \in D_n} \vec{d}_j$$

wobei α, β, γ Abstimmungskonstanten sind. Rochio hat $\alpha = 1$ gegeben.

Formel 22

$$sim(d_j, q) \propto \sum_{i=1}^{t} w_{i,q} w_{i,j} F_{i,j,q}$$

Formel 23

$$F_{i,j,q} = (C + idf_i)\overline{f}_{i,j}$$

$$\overline{f}_{i,j} = K + (1+K)\frac{f_{i,j}}{\max f_{i,j}}$$

wobei die Parameter *C* und *K* passend zu der Sammlung gesetzt werden.

Formel 24

$$F_{i,j,q} = \left(C + \log \frac{P(k_i \mid R)}{1 - P(k_i \mid R)} + \log \frac{1 - P(k_i \mid \overline{R})}{P(k_i \mid \overline{R})} \right) \overline{f}_{i,j}$$

$$P(k_i \mid R) = \frac{|D_{r,i}| + 0.5}{|D_r| + 1}$$

$$P(k_i \mid \overline{R}) = \frac{n_i - |D_{r,i}| + 0.5}{N - |D_r| + 1}$$

wobei $D_{r,i}$ die Menge der vom Nutzer als relevant gekennzeichneten Dokumente repräsentiert, in der die Dokumente die Ausdrücke k_i beinhalten.

Formel 25

$$Sim(q,K)=\prod_{k_i\in q}\left(\delta+\frac{\log\left(f\left(K,k_i\right)\times idf_K\right)}{\log n}\right)^{idf_i}$$

wobei $f\left(K,k_i\right)$ die Korrelationsfunktion zwischen dem Konzept *K* und den Anfrageausdrücke k_i ist. Diese Funktion wird folgendermaßen definiert:

$$f\left(K,k_i\right)=\sum_i pf_{i,j}\times pf_{K,j}$$

wobei $pf_{i,j}$ die Häufigkeit der entstehenden Ausdrücke k_i in der *j*-ten Passage von *n* Passagen ist. Wie $pf_{i,j}$ ist $pf_{K,j}$ die Häufigkeit des entstandenen Konzepts *K* in der *j*-ten Passage.

Die invertierte Dokumenthäufigkeit wird folgendermaßen berechnet:

$$idf_i=\max\left(1,\frac{\log_{10}N/np_i}{5}\right)$$

$$idf_K=\max\left(1,\frac{\log_{10}N/np_K}{5}\right)$$

wobei N die Anzahl der Passagen in der Sammlung ist,

np_i die Anzahl der die Ausdrücke k_i beinhaltenden Passagen ist,

np_K die Anzahl der das Konzept *K* beinhaltenden Passagen ist.

Der δ-Wert wird klein definiert, normalerweise beträgt er ca. 0,1.

Formel 26

$$itf_j=\log\frac{t}{t_j}$$

wobei *t* die Anzahl der unterschiedlichen Indexe in der Sammlung ist.

Formel 27

$$w_{i,j} = \frac{\left(0{,}5 + 0{,}5 \cdot \frac{f_{i,j}}{\max_l\left(f_{i,j}\right)}\right) \cdot itf_j}{\sqrt{\sum_{l=1}^{N}\left(0{,}5 + 0{,}5 \cdot \frac{f_{i,j}}{\max_l\left(f_{i,l}\right)}\right)^2 \cdot itf_l{}^2}}$$

wobei $f_{i,j}$ die Häufigkeit des Indexes k_i im Dokument d_j ist,

$\max_l\left(f_{i,j}\right)$ der maximale Wert der Häufigkeit $f_{i,j}$ in der Sammlung ist.

Formel 28

$$K_{u,v} = \vec{k}_u \cdot \vec{k}_v = \sum_{\forall d_j} w_{u,j} \cdot w_{v,j}$$

wobei $\vec{k}_u = \left(w_{u,1}, w_{u,2}, ..., w_{u,N}\right)$

Formel 29

$$\vec{q} = \sum_{k_i \in q} w_{i,q} \vec{k}_i$$

wobei das Gewicht $w_{i,q}$ wie in Formel 27 berechnet wird.

Formel 30

$$Sim\left(q, k_v\right) = \vec{q} \cdot \vec{k}_v = \sum_{k_u \in Q} w_{u,q} \cdot w_{u,v}$$

Formel 31

$$w_{v,q'} = \frac{Sim\left(q, k_v\right)}{\sum_{k_u \in q} w_{u,q}}$$

Formel 32

$$Sim(q, d_j) \propto \sum_{k_v \in d_j} \sum_{k_u \in q} w_{v,j} \cdot w_{u,q} \cdot K_{u,v}$$

Formel 33

$$n_{ab} = \sum_{i=1}^{n} \sum_{j=i-\delta}^{j=i+\delta} ind_{ab}(i, j)$$

wobei n die Anzahl der Wörter im Korpus ist,

δ die Größe des Fensters ist,

$$ind(i, j) = \begin{cases} 1 & falls\ w_i = a \wedge w_j = b \wedge doc(i, j) = 1 \\ 0 & sonst \end{cases}$$

$$doc(i, j) = \begin{cases} 1 & falls\ i\ und\ \ j\ im\ selben\ Dokument\ liegen \\ 0 & sonst \end{cases}$$

w_i das Wort an der Stelle i ist, $i=1,...,n$

Formel 34

$$ass(a,b) = \frac{\max(n_{ab} - \kappa \cdot n_a \cdot n_b, 0)}{(n_a + n_b)/2}$$

wobei κ eine Konstante abhängig vom Korpus ist.

Formel 35

$$sim(a,b) = \frac{\sum_{c \in A} ass(a,c) \cdot ass(c,b)}{\left[\sum_{c \in A} ass(a,c)^2\right]^{1/2} \left[\sum_{c \in A} ass(c,b)^2\right]^{1/2}}$$

wobei A die Menge der Wörter von der assoziierten erste Ordnung ist.

8 LITERATURVERZEICHNIS

[ACKE00] M. Ackermann: Statistische Korpusanalyse zum Extrahieren von semantischen Wortrelationen. *Dissertation*, Universität Hildesheim, 2000

[ACP01] M. Agosti, F. Crestani, and G. Pasi (Eds.): *Lectures on Information Retrieval*, Springer-Verlag, Germany, 2001

[AMRI02] G. Amati, C.J. van Rijsbergen: Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. In: *ACM Transactions on Information Systems(TOIS),* Band 20(4), 357-389, 2002

[BALL00] L.A. Ballesteros: Cross-Language Retrieval via Transitive Translation. In: Croft, W.B., *Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval*, Kluwer Academic Publishers, U.S.A., 2000

[BEAZ99] R. Beaza-Yates: *Modern Information Retrieval*, Addison-Wesley, 1999

[BENT06] H-J. Bentz: Suchen und Problemlösen in Komplexer Umgebung. In: *Perspectives on Cognition: A Festschrift for Manfred Wettler* (Hrg. von Rapp, R.; Sedlmeier, P.). Pabst Science Publishers, Lengerich, 2006

[BKSK99] M. Braschler, M. Kan, P. Schäuble und J. L. Klavans: The Eorospider Retrieval System and the TREC-8 Cross-Language Track. In: *Proceedings of TREC-8*, Gaithersburg, Maryland, USA, Nov. 1999

[BNX98] P. Buitelaar, K. Netter und F. Xu: Integrating Different Strategies for Cross-Language Information Retrieval in the MIETTA Project. In: *Proceedings of the 14th Twente Workshop on Language Technology (TWLT 14).* Language Technology in Multimedia Information Retrieval, December 7-8, Enschede, Niederlande, 9-17, 1998

[BOBU01] P. Bouillon und F. Busa: *The Language of Word Meaning*, Cambridge University Press, USA, 2001

[BORI00] L. Borin, Something Borrowed und Something Blue: Rule-Based Combination of POS Taggers. In: *Second International Conference on Language Resources and Evaluation. Proceedings*, Band I, Athens, Greece, 31 May–2 June, 21-26, 2000

[CCCS02] C. Chua, L. Cao, K. Cousins und D. W. Straub: Measuring Researcher-Production in Information Systems. In: *Journal of the Association for Information Systems*, Band 3, 145-215, 2002

[CHOE93] G. Chartrand und O.R. Oellermann: *Applied and Algorithmic Graph Theory*, McGraw-Hill, USA, 1993

[CHSO01] P. Charoenpornsawat, and V. Sornlertlamvanich: Automatic Sentence Break Disambiguation for Thai, In: *Proceedings of ICCPOL2001*, Korea, 231-235, May 2001

[CSI97] T. Charoenporn, V. Sornlertlamvanich und H. Isahara: Building a Large Thai Text Corpus -Part-Of-Speech Tagged Corpus ORCHID-. In: *Proceedings of the Natural Language Processing Pacific Rim Symposium 1997*, Phuket, Thailand, 509-512, 1997

[DLL96] S. T. Dumais, T. K. Landauer und M. L. Littman: Automatic cross-linguistic information retrieval using Latent Semantic Indexing. In *SIGIR'96 - Workshop on Cross-Linguistic Information Retrieval*, 16-23, August 1996

[ERIC95] B. Eric: Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging, In: *Computational Linguistics*, Band 21(4), December 1995, 543-565

[FLUH04] C. Fluhr, Multilinguality: http://www.LT-World.org/HLT_Survey/ltw-chapter8-5.pdf , Stand 2004

[FUWE02] N. Fuhr und G. Weikum: Classification and Intelligent Search on Information in XML. In: *IEEE Data Engineering Bulletin*, Band 25(1), 51–58, 2002

[GEJI99] F.C. Gey und H. Jiang: English-German Cross-Language Retrieval for the GIRT Collection – Exploiting a Multilingual Thesaurus, *Draft Proceedings of the Eighth Text REtrieval Conference (TREC-8),* National Institute for Standards and Technology, Washington, DC, November 17-19, 1999.

[GEY01] F.C. Gey: Research to improve Cross-Language Retrieval - Position Paper for CLEF. In C. Peters (Ed.): *CLEF 2000, LNCS 2069,* Springer-Verlag Berlin Heidelberg, 83-88, 2001

[GKP02] F. Gey, N. Kando und C. Peters: Cross Language Information Retrieval: a Research Roadmap. *Summary of a Workshop at SIGIR-2002: 22nd International Conference On Research And Development in Information Retrieval*, Tampere Finland, August 15, 2002

[GNXZZH98] J. Gao, J.Y. Nie, E. Xun, J. Zhang, M. Zhou und C. Huang: Improving Query Translation for Cross-Language Information Retrieval using Statistical Models. In: *Annual ACM Conference on Research and Development in Information Retrieval*, Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, New Orleans, Louisiana, United States, 96 – 104, 1998

[GRWE02] F.A. Grootjen und Th.P. van der Weide: Conceptual Relevance Feedback. In: *Proceeding of the 2002 IEEE International Conference on Systems*, Man and Cybernetics, (NLPKE 2002),Tunis, October 2002

[GRWE04] F.A. Grootjen und Th.P. van der Weide: Conceptual Query Expansion. *Technical Report NIII-R0406*, Nijmegen Institute for Information and Computing Sciences, University of Nijmegen, Nijmegen, Niederlande, 2004

[GAWI05] B. Ganter und R. Wille: *Formal Concept Analysis - Foundations and Applications*, Springer-Verlag GmbH, 2005

[HAGS96] M. Hagström: Textrecherche in großen Datenmengen auf der Basis spärlich codierter Assoziationsmatrizen. *Dissertation*, Universität Hildesheim 1996

[HEIT94] M. Heitland: Einsatz der SpaCAM-Technik für ausgewählte Grundaufgabe der Information. *Dissertation*, Universität Hildesheim 1994

[HOJO85] R.A. Horn und C.A. Johnson: *Matrix Analysis*, Cambridge University Press, USA, 1985

[IWSH00] Ł. M. Iwaniska und S.C. Shapiro: *Natural Language Processing and Knowledge Representation*, Menlo Park, CA/Cambridge, MA: AAAI Press/MIT Press., 2000

[JUMA00] D. Jurafsky und J.H. Martin: *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*, Upper Saddle River, NJ : Prentice Hall, 2000

[JARU01] C. Jaruskulchai: Dictionary-based Thai CLIR: Experimental Survey of Thai CLIR. In: *Lecture Notes In Computer Science Band 2406 - Revised Papers from the Second Work-*

shop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems-, Springer-Verlag, London, UK, 209 - 218, 2001

[KOEH05] P. Koehn: Europarl: A Multilingual Corpus for Evaluation of Machine Translation, http://www.isi.edu/~koehn/publications/europarl/ , In der Vorbereitung, Stand 2005.

[KOWA97] G. Kowalski: *Information Retrieval Systems – Theory and Implementation*, Kluwer Academic Publishers, 1997

[LALI90] T. K. Landauer und M.L. Littman: Fully automatic cross-language document retrieval using latent semantic indexing. In: *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research,* UW Centre for the New OED and Text Research, Waterloo Ontario, 31–38, October 1990

[LEZI00] W. Lezius: Morphy – German Morphology Part-of-Speech Tagging and Application. In: *Proceedings of the 9th EURALEX International Congress,* Stuttgart, Germany, 619-623, 2000

[LEZI94] W. Lezius: Morphologiesystem : Morphy. In R. Hausser, editor, *Linguistische Verifikation. Dokumentation zur ersten Morpholympics 1994*, Niemeyer, 25-35, 1994

[MASC02] C.D. Manning und H. Schütze: *Foundations of Statistical Natural Language Processing*, MIT press fifth printing, Cambridge, Massachusetts, London, England, 2002

[MCK97] S. Meknavin, P. Charoenpornsawat und B. Kijsirikul: Feature-based Thai Word Segmentation. In: *Proceedings of the Natural Language Processing Pacific Rim Symposium 1997(NLPRS'97),* Phuket, Thailand, 41-46, 1997

[MFKP99] H. Masuichi, R. Flournoy, S. Kaufmann und S. Peters: Query Translation Method for Cross Language Information Retrieval. In *Proceedings of the Workshop on Machine Translation for Cross Language Information Retrieval*, MT Summit VII, Singapore, 30-34, September 1999

[MISO00] P. Mittrapiyanuruk, und V. Sornlertlamvanich: The Automatic Thai Sentence Extraction, *The fourth Symposium on Natural Language Processing 2000*, Chiang Mai, Thailand, 23-28, May 2000

[MLG00] M. Montes-y-Gómez, A. López-López und A. F. Gelbukh: Information Retrieval with Conceptual Graph Matching**.** *In: Lecture Notes in Computer Science*, Band 1873, Proceedings of the 11th International Conference on Database and Expert Systems Applications, Springer-Verlag, 312 – 321, 2000

[MMI02] M. Murata, Q. Ma und H. Isahara: Comparison of Three Machine-Learning Methods for Thai Part-of-Speech Tagging, In: *ACM Transactions on Asian Language Information Processing (TALIP),* Band 1(2), 145 – 158, June 2002

[NABE05] S. Na nhongkai und H-J. Bentz: Bilinguale Suche mittels Konzeptnetzen. In: T. Mandl und C. Womser-Hacker (Hrsg.), Effektive Information Retrieval Verfahren in der Praxis: *Ausgewählte und erweiterte Beiträge des Vierten Hildesheimer Evaluierungs- und Retrievalworkshop (HIER 2005)* Hildesheim, 20. Juli 2005. Konstanz: Universitätsverlag [Reihe Schriften zur Informationswissenschaft 45], 2005.

[OADO96] D.W. Oard und B. J. Dorr: Evaluate Cross-Language Text Filtering Effectiveness. In: *Proceedings of the Cross-Linguistic Multilingual Information Retrieval Workshop*, ACM SIGIR Conference, Zurich, 8-14, August 1996

[OADO98] D.W. Oard und B. J. Dorr: Evaluate Resources for Query Translation in Cross-Language Information Retrieval. In : *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain, 759 – 763, 1998

[OARD97a] D.W. Oard: Alternative Approaches for Cross-Language Text Retrieval, In: *AAAI Symposium on Cross-Language Text and Speech Retrieval. American Association for Artificial Intelligence*, March 1997

[OARD97b] D.W. Oard: Cross-Language Text Retrieval Research in the USA, 3rd DELOS Workshop; Cross-Language Information Retrieval, number 97-W003 in: *Ercim Workshop Proceedings*. European Research Consortium for Informatics and Mathematics, March 1997

[PAZI99] M.T. Pazienza: *Information Extraction: Towards Scalable, Adaptable Systems*, Springer-Verlag, Germany, 1999

[PETE01] C. Peters: *Cross Language Information Retrieval and Evaluation: revised Papers / Workshop of the Cross Language Evaluation Forum, CLEF 2000*, Lisbon, Portugal, September 21-22, 2000, Springer Verlag, 2001

[PSC00] T. Potipiti, V. Sornlertlamvanich und P. Charoenpornsawat: Towards Building a Corpus-Based Dictionary for Non-Word-Boundary Languages. In: *Workshop on Terminology Resources and Computation, Workshop Proceedings of the Second Internationl Conference on Language Resources and Evaluation (LREC2000)*, Athens, Greece, 82-86, May 2000

[PSC00] T. Potipiti, V. Sornlertlamvanich und P. Charoenpornsawat: Automatic Corpus-Based Thai Word Extraction. In: *The Fourth Symposium on Natural Language Processing (SNLP2000),* Chiang Mai, Thailand, 176-181, May 2000

[RAPP99] R. Rapp: Automatic Identification of Word Translations from Unrelated English and German Corpora. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, College Park, Maryland, 519-526, 1999

[RAWI00] M. Ransburg, and K. Wiggisser: Elektronische Lexika, http://www.unfolded.com/writings/technical-papers/data/el/el.pdf, Stand 2000

[SCHM94] H. Schmidt: Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *International Conference on New Methods in Language Processing*, Manchester, UK, 1994

[SCHM95] H. Schmidt: Improvements in Part-of-Speech Tagging with an Application to German, In: *Proceedings of the 14th International Conference on Computational Linguistics*, Kyoto, Japan, 172-176, 1995

[SCI97] V. Sornlertlamvanich, T. Charoenporn und H. Isahara: ORCHID: Thai Part-Of-Speech Tagged Curpus. In: *Technical Report Orchid Corpus*, 1997

[SENT04] Whitepaper zur Suchmaschine SENTRAX Essence Extractor Engine, Manuskript Imbyte GmbH, Hildesheim 2004

[SKKRL02] H.C. Seo, S.B. Kim, B.I. Kim, H.C. Rim und S.Z. Lee: KUNLP System for NTCIR-3 English-Korean Cross-LanguageKUNLP System for NTCIR-3 English-Korean Cross-Language Information Retrieval. In K. Oyama, E. Ishida, & N. Kando (Eds.), *NTCIR Workshop3 Proceedings of the third NTCIR workshop on research in information retrieval, automatic text summarization and question answering*, Tokyo, Japan, NII, 2002

[SPWM00] V. Sornlertlamvanich, T. Potipiti, C. Wutiwiwatchai und P. Mittrapiyanuruk: The State of Art in Thai Language Processing. In: *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong, 597-598, October 2000

[STRA99] T. Strazlkowski: *Natural Language Information Retrieval*, Kluwer Academic Publishers, Niederlande, 1999

[SUPR99] P. Suwanvisat und S. Prasitjutrakul: Transliterated Word Encoding and Retrieval Algorithms for Thai-English Cross-Language Retrieval. In: *Proceedings of the National Computer Science and Engineering Conference 1999*, Bangkok, Thailand, Dec.16–17, 1999

[SUSM01] R. Sukhahuta und D. Smith: Information Extraction Strategies for Thai Documents. In: *International Journal of Computer Processing of Oriental Languages (IJCPOL)*, Band 14(2), 153-172, 2001

[TAKK97] J. Takkinen: CAFE: Towards a Conceptual Model for Information Management in Electronic Mail. *Doctor Thesis*, School of Engineering at Linköping University, 1997

[TCTS00] T. Treeramunkong, W. Chinnin, T. Tanhermhong und V. Sornlertlamvanich: Full Text Search for Thai Retrieval Information System. *The fourth Symposium on Natural Language Processing 2000*, 2000

[TSTC00] T. Treeramunkong, V. Sornlertlamvanich, T. Tanhermhong und W. Chinnan: Character cluster based Thai information retrieval. In: *Proceedings of the fifth international workshop on on Information retrieval with Asian languages*, Hong Kong, China, 75 - 80, 2000

[TUMA98] D. Tufis und O. Mason: Tagging Romanian texts: A Case Study for QTAG, a Language Independent Probabilistic Tagger. In: *Proceedings of the First International Conference on Language Resources & Evaluation (LREC),* Granada(Spain) , 28-30 May, 1998, 589-596, 1998

[WCD02] D. Widdows, S. Cederberg und B. Dorow: Visualisation Techniques for Analysis Meaning. In: *Fifth International Conference on Text, Speech and Dialogue (TSD5),* Brno, Czech Republic, 107-115, September 2002

[WEIP01] E. Weippl: Visualizing Content Based Relations in Texts. In: *ACM International Conference Proceeding Series Band 14 - Proceedings of the 2nd Australasian conference on User interface-,* Queensland, Australia, 34-41, 2001

[WIDO02] D. Widdows und B. Dorow: A Graph Model for Unsupervised Lexical Acquisition. *19th International Conference on Computational Linguistics*, Taipei, 1093-1099, August 2002

[XUCR96] J. Xu und W.B. Croft: Query Expansion Using Local and Global Dokument Analysis. In: *Proceedings of the 19th annual international ACM SIGIRconference on Research and development in information retrieval*, ACM Press, 4–11, 1996

[XWBJ01] L. Xiao, D. Wissmann, M. Brown und S. Jablonski: Hierarchical Concept Description and Learning for Information Extraction. In: *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium (NLPRS 2001)*, Tokyo, Japan, 27-30November 2001, 299-306, 2001

[ZAND06] M. Zander: Automatisierte Wortlistenerzeugung durch multiple Dokumentenreduzierung im bilingualen Kontext, *Diplomarbeit*, Institut für Mathematik und angewandte Information, Universität Hildesheim, 2006

LEBENSLAUF

14.09.1970	Geburt in Ratschaburi, Thailand
1975 – 1983	Besuch der Grundschule Anuban-Ratschaburi in Ratschaburi
1983 – 1986	Besuch der Mittelschule Benjama-Raschutit-Ratschaburi in Ratschaburi
1986 – 1989	Besuch der Oberschule Sribunjanon in Nonthaburi mit dem Stipendium „The Development and Promotion of Science and Technology Talents Project“
1989 – 1993	Studium für “Bachelor degree of science B.Sc. (Mathematics)” mit dem Hauptfach Mathematik und dem Nebenfach Informatik an der Kasetsart University, Bangkok, Thailand, mit dem Stipendium „The Development and Promotion of Science and Technology Talents Project“(1989-1991)
1993 – 1999	Studium für “Master degree of science M.Sc. (Mathematics)” an der Kasetsart University
1993 – 1994	Anstellung als Privatdozent für Mathematik an der Mahanakon University of Technology, Lehrtätigkeit in Analysis und Mathematik-Software
1994 – 1995	Mitarbeiter in der Arbeitgruppe “Parallel computing” bei Assoc. Prof. Dr. Royol Jittradon an der Kasetsart University
1995 – 1999	Mitarbeiter an der Labor “High Performance Computing Center (HPCC)“, National Electronics and Computer Technology Center (NECTEC)
ab 1996	Anstellung an der Kasetsart University als Dozent im Fachbereich Mathematik mit der Lehrtätigkeit in Analysis und Numerische Methoden
1997 – 1999	Anstellung als externer Dozent an der privaten Universität St.John, Unterricht in Grundlagenmathematik und Analysis
1999 – 2000	Anstellung als externer Dozent an der Kasetsart University (Srirascha Campus), Unterricht in Linieare Algebra
2001 – 2006	Promotionsstudium an der Universität Hildesheim, Institut für Mathematik und Angewandte Informatik.